NÁVRH PROJEKTU DO VEŘEJNÉ SOUTĚŽE VE VÝZKUMU A VÝVOJI

NÁRODNÍ PROGRAM VÝZKUMU II

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Evideční číslo projektu - přidělí poskytovatel

Akronym (Podací číslo projektu)

COT-SEWing

Název projektu

Prostředky tvorby komplexní báze znalostí pro komunikaci se sémantickým webem v přirozeném jazyce

PROGRAM

INFORMAČNÍ TECHNOLOGIE PRO ZNALOSTNÍ SPOLEČNOST

Tematická oblast

Eliminace jazykových bariér prostředky informačních technologií

Téma

Nové postupy nebo návrhy zařízení umožňující vytvoření informačního základu - komplexní báze znalostí pro budování různých lingvistických aplikací

Cíl projektu

Vytvoření souboru nových prostředků – lingvistických aplikací, založených na metodách umělé inteligence, umožňujících komunikaci s webovým prostředím na kvalitativně vyšší úrovni.

Doba řešení

2006 - 31.12.2010

Uchazeči - řešitelská pracoviště - řešitelský tým (odpovědnost za řešení)

Západočeská univerzita v Plzni (příjemce - koordinátor) - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky (řešitelské pracoviště)
Ježek Karel doc. Ing. CSc. (řešitel koordinátor)

Masarykova univerzita (spolupříjemce) - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka (řešitelské pracoviště)
Pala Karel doc. PhDr. CSc. (spoluřešitel)

1. IDENTIFIKAČNÍ ÚDAJE PROJEKTU

Akronym projektu

COT-SEWing

Název projektu

Prostředky tvorby komplexní báze znalostí pro komunikaci se sémantickým webem v přirozeném jazyce

Anotace projektu

Vypracovat komplexní systém metodických a implementačních nástrojů na bázi inteligentních agentů pro vytváření uživatelsky přívětivých rozhraní k sémantickému webu umožňující, aby komunikace s uživatelem probíhala v přirozeném jazyce a též zpracovávaná data měla převážně charakter vět, resp. promluv, přirozeného jazyka. Dalším cílem pak je ověření funkčních vlastností navržených nástrojů na vhodně zvolené aplikaci.

Soutěž

NÁRODNÍ PROGRAM VÝZKUMU II - Ministerstvo školství, mládeže a tělovýchovy [VSMSMT6NPV2]

Program

INFORMAČNÍ TECHNOLOGIE PRO ZNALOSTNÍ SPOLEČNOST [2C]

Tematická oblast programu

Eliminace jazykových bariér prostředky informačních technologií [2C-1-5]

Téma projektu

Nové postupy nebo návrhy zařízení umožňující vytvoření informačního základu - komplexní báze znalostí pro budování různých lingvistických aplikací [2C-1-5-1]

2. PŘEDSTAVENÍ PROJEKTU

2.1. Představení řešení projektu

Prostředí Webu je dnes nejdůležitějším zdrojem informací. Jeho oblíbenost vychází i ze skutečnosti, že k tomuto médiu má přístup široké spektrum různě zaměřených uživatelů. Celosvětově se však stále výrazněji začínají projevovat negativní stránky spojené s často neomezenou dostupností Webu komukoliv, s virtuálně neomezovaným a tudíž rychle rostoucím množstvím v něm uchovávaných informací, s nekonzistencí forem používaných k vyjádření informací v počítačích s formou přirozeného jazyka apod. Nezanedbatelným problémem je také rychlost vyhledávání informací na webových stránkách, které představují typické distribuované prostředí, stále více se na webových stránkách objevují informace vyjádřené v některém z přirozených jazyků. Tím vzniká jedna z velkých informačních zábran širokého využití Webu – multijazykový charakter jeho obsahu. Smyslem navrhovaného projektu proto bude vytvoření komplexní báze programových prostředků, které umožní odstranění některých dnes typických zábran a některé další výrazně omezí, resp. běžnému uživateli jejich překonání výrazně usnadní.
Naším příspěvkem k řešení uvedených problémů bude vytvoření souboru nových nástrojů, algoritmů, metod a postupů, které dovolí uživateli vstřícným způsobem kvalitativně dokonalejší využívání Webu zahrnující:
- dotazování v přirozeném jazyce včetně pokládání dotazů běžnou mluvenou řečí,
- umožnění dalšího zpřesnění výsledku dotazu na základě rozpoznání oblastí zájmu uživatele,
- extrakci relevantní informace uložené v databázích obsahujících údaje a informace reprezentované v přirozeném jazyce, a to i hovorovou formou (řečové databáze),
- nástroje pro automatickou tvorbu sémantického popisu vět, resp. řečových promluv, v omezených významových doménách,
- prostředky pro získání výsledku dotazu i v multijazykové podobě, resp. v některém ze zvolených jazyků,
- prostředky pro klasifikaci a filtraci webových stránek a dokumentů,
- prostředky pro vytváření anotací a sumarizaci rozsáhlých dokumentů a jejich kolekcí,
- získávání latentních informací a znalostí z webového prostředí,
- nástroje pro evaluaci a vyhledávání doménových expertů, expertních skupin, autorit apod.

Získávání znalostí z databází přirozeného jazyka vyžaduje extrakci významu z analyzovaného textu, resp. rozpoznané promluvy, přičemž v případě mluveného projevu je nejprve nezbytné jeho spolehlivé rozpoznání (ASR), tedy jeho převedení z podoby řečového signálu do psané formy. Poté pro korektní identifikaci hledaných znalostí následuje sémantická analýza. V neposlední řadě je třeba řešit problematiku efektivní organizace extrahovaných dat a účinné metody jejich hledání. Pokud se týká úrovně poznání ve světě, v Evropě a v ČR zvláště, pak v současné době byly dopracovány ke komerčnímu využití ASR aplikace v oblastech diktování (nahrazení klávesnice – Nuance, IBM, Microsoft,...), v telefonních aplikacích (rezervace, vyhledávání v tel. seznamu, ...), v automatické transkripci (např. zpravodajských relací) a pomoci zrakově a tělesně postiženým. Nejvíce ASR i dalších produktů existuje pro angličtinu (vznikly většinou v USA), nejpokročilejší systémy rozpoznávající plynulou řeč dokážou pracovat se slovníky o velikosti až 300 tis. slov, resp. slovních forem, v reálném čase, při dosažení 98-99% úspěšnosti rozpoznávání (v tichém prostředí). Evropa se sice řídí celosvětovými trendy, ale je zde dosti patrné zaměření na jednotlivé národní jazyky. Mezi významná evropská pracoviště patří pracoviště fy Siemens, (výzkumné centrum v Mnichově), výzkumné centrum Daimler-Chrysler v Ulmu, DFKI Saarbrücken, Univerzita Kaiserslautern, IDIAP Martigny, LIMSI–CNRS Orsay, IBM Böblingen. Vznikají projekty anotací národních jazykových korpusů a tvorba ontologií pro jednotlivé národní jazyky. Velmi významným projektem v Evropě byl mimo jiné projekt Verbmobil (pro automatický překlad). Významný je tím, že pro dosažení kvalitního výstupu používá propracovaný systém sémantické analýzy. Dále byl v Evropě v devadesátých letech vyvinut koncept dialogového systému v rámci projektu Sundial, který posloužil k vytvoření řady jednoduchých aplikací dialogových systémů. V ČR byl vyvinut systém s pravděpodobně největším počtem úspěšně rozpoznávaných slovních forem na TU v Liberci (tým prof. Nouzy).
Vlivem evropské integrace, sbližování států, stíráním kulturních hranic aj. stále více vystupuje do popředí nutnost řešení vícejazyčnosti projevů, záznamů a dokumentů jako velmi aktuální. Mnohé současné systémy pro zpracování textů pracující s informacemi uloženými na Webu předpokládají pouze jednojazyčné prostředí a svou funkci tomu mají uzpůsobenu. Možnost uložení a zpracování vícejazyčných dokumentů buď vůbec neřeší, nebo pouze okrajově. Možno konstatovat, že ještě výraznější monolingualitu lze nalézt u audiovizuálně zaznamenaných dokumentů, navíc, zejména v rámci EU, jsou vytvořeny v různých jazycích. Jednotlivé dokumenty, ať už internetové, nebo uložené v některé digitální knihovně (knihovny, sbírky zákonů, právní předpisy, vědecké články, archivy rozhlasových a televizních pořadů, záznamy jednání parlamentů apod.) je ovšem nutné pro lepší orientaci uživatele třídit, filtrovat, či umožnit jejich prohledávání. Ideou projektu je umožnit jejich zpracování tak, aby byla informace z cizích jazyků vzájemně propojena a pro uživatele byla prezentována ve formě zahrnující všechny jazyky. Vzniká tak potřeba vyhledávat informace nezávisle na použitém jazyku, k čemuž je zapotřebí nalézt takový formální aparát pro reprezentaci významů, který nebude na existujících buď vůbec, nebo jen velmi málo závislý.
První myšlenkou zabývající se tímto problémem je myšlenka vytvoření sémantického webu. Ačkoli tato myšlenka je dnes velmi aktuální, k jejímu naplnění se zatím nepodařilo mnoho přispět. Jedním z důvodů je právě absence nezávislé formální reprezentace sémantiky a tu bude možné navrhnout až po velmi důkladné

sémantické analýze obrovského množství dokumentů uložených, resp. přístupných, prostřednictvím webu. V oblasti počítačového zpracování sémantiky je v současné době řešena především otázka efektivnějšího vyhledávání na Internetu. Další využití sémantické analýzy a nezávislé reprezentace sémantiky lze nalézt v inteligentních dialogových systémech, kde korektní porozumění uživatelově promluvě je klíčovým momentem funkce celého systému, a dále pro systémy automatického překladu (machine translation), kde teprve přechod na vyšší abstraktní rovinu popisu promluvy / věty zajistí kvalitnější a přirozenější překlad.
Originalita řešení spočívá v plně automatizovaném, na objektivních a ověřitelných skutečnostech založeném postupu, který bude analyzovat data dostupná prostřednictvím webu komukoliv. Vyjdeme z modelů sociálních sítí umožňujících určit objekty, které mají největší vliv na své okolí a mohou tedy být označeny za nejvlivnější, nejdůležitější, nejpopulárnější, nejkvalitnější a jinými synonymy podle typu dané sociální sítě. Klasickým případem je vědecká obec, v jejímž středu stojí lidé, jejichž práce je významná a užitečná pro ostatní a je tedy často využívána (citována). Sociální sítě jsou modelovány grafy a existuje mnoho algoritmů pro získávání různých informací, mj. i určování nejvýznamnějších objektů sítě. Sociální sítě jsou mnoha druhů a těmito objekty mohou být jak vědci, tak přímo jejich publikace nebo např. i webové stránky.
V rámci navrhovaného projektu bude vytvořena báze nástrojů a prostředků umožňujících efektivní přístup k informacím uloženým v databázích (a to i v databázích s informacemi reprezentovanými v přirozeném jazyce) přístupných přes webové stránky, vyhledávání v nich, extrakci zvolených informací apod. na základě sémantické reprezentace a jejich poskytování uživateli přirozenou formou. Charakter jedinečnosti řešení je zdůrazněn i zahrnutím do projektu výzkumu a vývoje prostředků pro české prostředí. K dosažení cílů bude nutno užít různých postupů, od automatizovaného procházení webu, přes extrakci informací z textu statistickými metodami (např. Markovovými modely), klasifikaci dokumentů lingvistickou analýzou, až po optimalizaci SQL dotazů nad databází. Cílová řešení budou tvořit bázi technické infrastruktury pro práci se znalostmi v prostředí Internetu, se zřejmým širokým uplatněním v oblasti výzkumu, školství, průmyslového vývoje, komerce, státní správy a organizací, které budou stále častěji přicházet do styku s cizojazyčnými dokumenty.

2.2. Garanti

- **Ježek Karel doc. Ing. CSc.** - Západočeská univerzita v Plzni (49777513)
- **Materna Pavel prof. RNDr. CSc.** - Masarykova univerzita Brno (00216224)
- **Matoušek Václav prof. Ing. CSc.** - Západočeská univerzita v Plzni (49777513)
- **Pala Karel doc. RNDr. CSc.** - Masarykova univerzita Brno (00216224)
- **Šafařík Jiří prof. Ing. CSc.** - Západočeská univerzita v Plzni (49777513)
- **Vávra František doc. Ing. CSc.** - Západočeská univerzita v Plzni (49777513)

3. RÁMEC PROJEKTU

3.1. POSLÁNÍ PROJEKTU

3.1.1. Definice účelu projektu

Posláním projektu bude návrh modelů a praktické ověření technik a postupů, které by byly schopny řešit problémy související s novými, perspektivními způsoby využití Webu, dovolujícími prostřednictvím metod počítačové lingvistiky extrahovat z jeho obsahu nejen informace, ale i skryté znalosti. Tyto by v dalším měly být ukládány do komplexní báze znalostí, která umožní odstranění některých typických problémů v přístupu k informacím uloženým na webu a běžnému uživateli přístup významně usnadní. Nepominutelnou součástí poslání projektu bude vytvoření souboru lingvistických nástrojů a programových produktů pro práci se sémantickým webem, a to nejen v českém jazyce.
Projekt je přihlašován do programu 2C, „Informační technologie pro znalostní společnost“, tématická oblast 2C-1-5 „Eliminace jazykových bariér prostředky informačních technologií“, oblast poslání 2C-1-5-1 „Nové postupy nebo návrhy zařízení umožňující vytvoření informačního základu – komplexní báze znalostí pro budování různých lingvistických aplikací“.
Moderní informační a komunikační technologie (Internet, mobilní sítě, dialogové systémy) poskytují nepřeberné množství informací. Na rozdíl od středověku, kdy byl jedinec schopen pojmout všechno dostupné vědění, je dnes orientace v tomto množství informace pro člověka nemožná bez moderních technologií. Příkladem takové technologie je textový vyhledávací stroj Google, který je dnes první stránkou mnoha webových prohlížečů na celém světě. K disposici jsou již v současné době mnohem bohatší zdroje dat obsahující mnohdy také audio a video záznamy – s postupným zlevňováním rychlého připojení k Internetu se tato data stávají zcela běžnou komoditou. Orientace v těchto modalitách přináší nové výzvy: algoritmy založené na indexování textu jsou zde nepoužitelné bez podstatné adaptace a nové modality si žádají zcela nové přístupy. Druhou podstatnou oblastí je multimodální komunikace s počítačem; v moderních systémech pro komunikaci s uživatelem (dialogové systémy, informační kiosky, reklama) je textový výstup jen základní modalitou, která je v mnoha aplikacích překonána. Pevně dané dialogy jsou nahrazovány inteligentním řízením, tzv. mixed initiative approach, které se adaptuje na uživatele, kontext svého použití a využívá reprezentace znalostí reálného světa. V rámci projektu budou proto vytvořeny metodické, formální a programové nástroje, které budou využitelné pro aplikace zpracování přirozeného jazyka a v oboru lingvistiky. Účelem je umožnit zpracování rozsáhlých kolekcí informací reprezentovaných jak běžným způsobem, tj. převážně texty, tak i soubory obsahujícími řečová data. Díky multilinguálnímu zpracování pak bude umožněno jedním dotazem získat všechny relevantní informace.

3.1.2. Očekávané přínosy projektu

Očekávané přínosu projektu lze shrnout do následujících bodů:
1. Posun state-of-the-art v jednotlivých technologiích jako jsou rozpoznávání řeči, počítačová lingvistika, audiovizuální rozpoznávání, analýza textů i souvislé řeči či vyhledávání v textových a řečových databázích. Uživateli výsledků této kategorie budou akademické laboratoře, které je využijí k dalšímu vývoji, i firmy, které mohou vyvinuté algoritmy uplatnit při vlastním komerčním výzkumu a vývoji.
2. Kvalitativní posun při zpracování českého jazyka, kdy standardní a ve světě používané algoritmy není možné použít díky ohebnosti a obrovské slovní zásobě českého jazyka. Uživateli těchto výsledků budou akademické laboratoře i firmy, které budou zabudovávat algoritmy pro zpracování jazyka do svých aplikací. V některých případech, jako jsou jazykové slovníky, také široká veřejnost.
3. Integrace uvedených technologií do funkčních systémů a objektivní (databázové) a subjektivní (uživatelské testy) vyhodnocení funkčnosti takových systémů. Uživateli budou především firmy, které získají ucelená řešení vhodná k rychlému zabudování do vlastních aplikací.
4. Maximální zjednodušení přenosu poznatků z akademické sféry do komerčního využití. Uživateli těchto výsledků budou kromě účastníků projektu všechny akademické instituce a rovněž komerční sféra; výsledky by měly přispět k přemostění „údolí smrti“ mezi akademickým výzkumem a komerční sférou vytvořením souboru přímo využitelných nástrojů.

3.1.3. Způsob ověření dosažených přínosů

Jednotlivé přínosy projektu budou ověřovány následujícím způsobem:

1. Posun state-of-the-art v technologiích přístupu k datům různého typu uloženým na webu bude ověřován standardními akademickými metrikami jako množstvím publikací, počtem citací, vyzvaných přednášek, atd. Specifikem pro navrhovanou oblast data jsou evaluační kampaně, kdy se několik výzkumných skupin utká o nejlepší výsledek (např. rozpoznávání, sumarizace textu, extrakce významu z daného textu) na stejných datech.
2. Pokrok při zpracování českého jazyka bude ověřován podobnými metrikami jako v bodě 1., důraz bude kladen na použitelnost vyvinutých řešení reálnými českými uživateli bez speciálních lingvistických znalostí.

Měřítkem zde bude mj. počet aplikací pro zpracování českého jazyka dostupných na trhu v průběhu a po ukončení projektu.
3. Měřítka kvality integrace budou dvě – integrace jednotlivých přístupů v rámci extrakce znalostí z dat různých modalit bude hodnocena objektivními či subjektivními kritérii, která vyhodnotí úspěšnost takového systému oproti systémům využívajícím pouze jednu modalitu (např. extrakce znalostí z řečové databáze versus pouze extrakce z textu). Kvalita schopnosti integrace jednotlivých komponent do celkového systému bude posuzována jak samotnými vývojáři, tak uživateli výstupů např. ve firmách; reakce těchto uživatelů budou pečlivě analyzovány.
4. Kvalita přenosu technologií mezi akademickou a komerční sférou bude měřena počtem vyvinutých demonstrátorů, počtem nově uzavřených nebo obnovených kontraktů, intenzitou komunikace na odborných veletrzích a počty ohlasů na činnost řešitelů v médiích se širokým dopadem (nikoliv odborné časopisy, ale noviny, televize, internetové informační portály, atd.).

Přínosy plynoucí z projektu budou dále ověřovány jak přímými, tak nepřímými metodami. K přímým metodám řadíme:
- testování navržených a implementovaných metod testovacím korpusem,
- použití metody přímého srovnání navrženého systému s existujícími obdobnými systémy,
- využití jak subjektivní metody porovnání systémů (ohodnocení výsledného řešení skupinou uživatelů – dotazníková forma), tak metody objektivní (např. F-míra).
V případě nepřímého způsobu ověření přínosů je plánována pravidelná prezentace výsledků v průběžných oponenturách, na lokálních a především na mezinárodních konferencích dedikovaných problematice řešené v projektu. Ucelené části řešení budou nabídnuty k publikaci impaktovaným časopisům.

3.1.4. Kritické předpoklady dosažení účelu projeku

Řešení projektu úzce navazuje na dosavadní činnost obou partnerů, jsou tedy vytvořeny dobré podmínky pro úspěšnou a bezproblémovou činnost vzniklého konsorcia. Rizikové faktory (RF) nicméně existují, následující body podávají jejich přehled a nástin řešení:

RF1: Omezení rozpočtu projektu během řešení v důsledku změny politiky vlády nebo MŠMT.
Řešení: Nutnost redukce zamýšlených cílů, případně nutnost hledat více zdrojů z českých a evropských grantových projektů a z eventuální spolupráce s partnery z komerční sféry. U obou partnerů není plánovaný projekt jediným zdrojem financování výzkumu, proto je toto řešení schůdné, byť nepříjemné. V krajním případě rozvázání pracovních poměrů s již najatými pracovníky, s negativním dopadem na množství a kvalitu dosažených výsledků.

RF2: Neschopnost jednoho z partnerů splnit dílčí úkol.
Řešení: Oba zúčastnění partneři se danou problematikou zabývají již po mnoho let, takže toto riziko je velmi nepravděpodobné. K mírnému ohrožení plánovaných výsledků by mohlo dojít snad jedině úmrtím nebo ze zdravotních důvodů některého ze zúčastněných řešitelů; obě pracoviště jsou však natolik dobře personálně vybavena, že by zřejmě nenastal problém s přesunem úkolů na jiného pracovníka. Přesun úkolů na druhého partnera není možný (každý ze zúčastněných partnerů řeší své specifické úkoly).

RF3: Zánik některé z laboratoří účastnících se na řešení projektu.
Řešení: Vzhledem ke stabilitě obou zúčastněných univerzitních laboratoří a relativně krátké době řešení projektu je toto riziko velmi nepravděpodobné.

RF4: Kvalita výsledků produkovaných navrženými postupy a metodikami..
Řešení: Srovnání se stávajícími systémy a s výsledky dosaženými dalšími výzkumnými skupinami by měla prokázat, že řešení navržená v rámci navrhovaného projektu poskytují v daném kontextu kvalitnější výstupní hodnoty. Neméně podstatným faktorem je i časová a paměťová složitost vyvinutých algoritmů, které samozřejmě určují praktickou použitelnost jednotlivých postupů a modelů.

3.2. CÍL PROJEKTU

3.2.1.Definice cíle projektu

3.2.1.1. Čeho (Co má být projektem dosaženo

Vytvoření souboru nových prostředků – lingvistických aplikací, založených na metodách umělé inteligence, umožňujících komunikaci s webovým prostředím na kvalitativně vyšší úrovni.

3.2.1.2. Do jakého data bude dosaženo cíle

31.12.2010

3.2.2. Výsledky projektu

Dokonalejší komunikace uživatele s webovým prostředím bude dosaženo nejen vývojem nových, dokonalejších modelů a metod vycházejících z klasických principů statistiky, lingvistiky a umělé inteligence, ale zejména zužitkováním sémantických informací z obsahu Webu. Vytvořením předpokládané báze znalostí a aplikací výše zmíněných metod a modelů bude dosaženo:
- zlepšení technik a výsledků vyhledávání na webových stránkách a v souvisejících databázích,
- uživatelsky příjemnějšího a komfortnějšího vyhledávání informací na webu,
- zpřístupnění webu širší skupině uživatelů, včetně uživatelů handicapovaných,
- komunikace s webem prostředky přirozeného jazyka,
- rozšíření nabídky komunikačních prostředků při práci s informacemi na webu,
- vytvoření nástrojů a metod syntaktické analýzy textů v přirozeném jazyce (češtině, popř. dalších jazycích); tyto budou základem pro aplikace, které pracují s volnými texty z pohledu jejich vnitřní struktury – typickými aplikacemi jsou strojový překlad, extrakce a identifikace termínů v textu či logická a sémantická analýza textu,
- vytvoření nástrojů pro logickou analýzu české věty za účelem reprezentace znalostí, tj. automatický překlad věty do zápisu ve formě logické formule, který zahrne algoritmy pro řešení anaforických vztahů (zájmen) a poskytne informatický základ pro automatické budování báze znalostí z textu,
- verifikace algoritmů pro vyvozování nových faktů založených na znalostech extrahovaných z volného textu pro vybrané domény – výsledky umožní inteligentní zpracování dotazu, jehož podklady jsou založené na textových datech, a určení odpovědi, která se nezakládá pouze na vyhledání vzorů v textu,
- vytvoření sémantických formalismů a jejich využití k reprezentaci významu vět/promluv s cílem dosažení dokonalejší komunikace s webem,
- využití struktury i sémantiky webových stránek k extrahování znalostí o jejich majiteli, např. evaluačních hledisek pro hodnocení školy, firmy apod.,
- lepší orientace ve vyhledaných informačních zdrojích.

Přínosem rovněž budou textové a strukturní korpusy (jako je korpus syntaktických stromů nebo korpus přepisu sémantické reprezentace vět) a thesaury pokrývající různé tématické oblasti, které je potřeba vytvořit k natrénování a ověřování algoritmů a které by mohly být využívány i pro následné výzkumné či výukové účely. Přitom propojení češtiny s ostatními jazyky zemí sdružených v EU na platformě www chápeme jako prioritní úkol projektu.

3.2.3. Forma zpracování a předání výsledků

Výsledky projektu budou popsány v závěrečné dokumentaci, veškeré navržené postupy budou podrobně popsány, výsledky experimentů analyzovány a zhodnoceny z hlediska jejich přínosu. O průběžných výsledcích budou informovat ročně zpracovávané technické zprávy a webové stránky projektu. Součástí zpráv bude i distribuční médium vytvořených aplikací a kopie souvisejících publikací. Závěrečná zpráva i průběžné zprávy budou oponovány.

Jedním z významných nástrojů pro průběžnou prezentaci výsledků práce na projektu bude organizace mezinárodní konference Text, Speech and Dialogue (TSD), kterou pravidelně od roku 1998 spolupořádají uchazeči FAV ZČU Plzeň a FI MU Brno. Konference TSD se za 9 let vyvinula v primární fórum pro výměnu zkušeností mezi výzkumníky z oblastí zpracování psaného i mluveného jazyka ze zemí původního východního bloku a jejich západních kolegů. Sborník konference TSD vychází pravidelně v nakladatelství Springer-Verlag v řadě Lecture Notes in Artificial Inteligence; poslední vydání sborníku vyšlo pod pořadovým číslem LNAI 3658 v září 2005.

Kvalita výsledných postupů a programů bude ověřena na několika úrovních. Vlastní měřitelné charakteristiky budou průběžně kontrolovány proti dostupným i vytvořeným korpusovým datům. Zjištěné údaje budou srovnány s výsledky odpovídajících projektů v ČR i v zahraničí a tato srovnání budou prezentována na národních a mezinárodních konferencích, kde budou posouzeny oponenty, editory a také všemi účastníky. V rámci projektu budou samozřejmě hodnocení práce součástí průběžných zpráv i závěrečné zprávy.

Vytvořené programové systémy budou k dispozici jako samostatné nástroje nebo programové knihovny. Jako

takové budou tvořit vývojový systém pokrývající podstatnou část zpracovávané tématické oblasti.

3.2.4. Kritické předpoklady dosažení cíle

Základním předpokladem k dosažení definovaných cílů je zachování a přiměřená podpora výzkumného týmu na pracovišti projektu. Kritické předpoklady dosažení účelu projektu týkající se organizace a financování jsou shrnuty v sekci 3.1.4 této přihlášky. Rizikové faktory ovlivňující vědeckou a technologickou část projektu a nástin jejich řešení jsou následující:

RF1: Nepotvrzení či neplatnost výzkumných hypotéz poskytujících základ další činnosti týmu.
Řešení: Projekt, resp. jeho cíle, nestojí na jedné výzkumné hypotéze, ale na teoretickém základu zpracování multimodálních dat. Pro činnost obou výzkumných skupin existuje určité množství experimentálních dat pro trénování a testování vytvářených systémů, jejich množina se bude v průběhu řešení projektu nadále intenzivně doplňovat. Jejich tvorba je činnost zdlouhavá a nákladná, avšak bezriziková. Z dostupných prací vyplývá, že větší množství dat vede vždy ke zlepšení vlastností systému.

RF2: V průběhu projektu přestane být o vytvářené přístupové technologie zájem a pracoviště účastnící se na řešení projektu se tak ocitnou bez reálné využitelnosti svých výsledků.
Řešení: Současným trendem je naopak příklon k ukládání množství dat a informací na běžných počítačových prostředcích, sílí propojování informačních technologií s rozhlasovým a televizním vysíláním, streamovanými médii a mobilními komunikacemi. Nové hardwarové prostředky budou vyžadovat nové technologie přístupu k datům, přičemž preferována bude komunikace v přirozeném jazyce, ať už psanou nebo mluvenou formou. Proto je toto riziko za dobu řešení projektu téměř nulové.

RF3: V průběhu projektu se vyskytne komerční software řešící problematiku srovnatelnou s předpokládanými výsledky projektu.
Řešení: Komerční řešení využívající přístup k datům na webu prostřednictvím přirozeného jazyka jsou dosud v plenkách a komerční sféra naopak aktivně vyhledává zajímavé práce z akademické sféry. Navíc řešení komunikace v přirozeném jazyce vyžaduje velmi rozsáhlé know-how, které běžné firmy zpravidla nemají. Proto je toto riziko minimální, očekáváme naopak velký zájem z komerční sféry.

3.3. DÍLČÍ CÍLE ŘEŠENÍ - přehled

1
Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím.
- 31.12.2007 - Aplikovaný výzkum s výjimkou průmyslového výzkumu (tzv. "neprůmyslový výzkum")

2
Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka.
- 31.12.2008 - Základní výzkum

3
Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce.
- 31.12.2009 - Aplikovaný výzkum s výjimkou průmyslového výzkumu (tzv. "neprůmyslový výzkum")

4
Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí.
- 31.12.2010 - Aplikovaný výzkum s výjimkou průmyslového výzkumu (tzv. "neprůmyslový výzkum")

3.3.x. DÍLČÍ CÍL ŘEŠENÍ PROJEKTU - "1" - podrobně

3.3.1. Definice dílčího cíle

3.3.1.1. Co má být dílčím cílem dosaženo

Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím.

3.3.1.2. Datum dosažení dílčího cíle

31.12.2007

3.3.1.4. Typ činnosti při řešení dílčího cíle

Aplikovaný výzkum s výjimkou průmyslového výzkumu (tzv. "neprůmyslový výzkum")

3.3.2. Výsledky dílčího cíle

a) Vytvoření uživatelského rozhraní pro hlasový vstup / příp. výstup, které bude použito pro komunikaci se sémantickým webem, a pro jeho podporu vytvoření robustního ASR systému pro inflexní jazyky. K tomu bude nutno vytvořit kvalitní korpus pro ASR a z něj extrahovat dostatečné množství trénovacích dat. V jednotlivých etapách bude v průběhu let 2006 – 2007 vytvořen:
- kvalitní audio-korpus pro natrénování systému ASR,
- korpus pro natrénování jazykových modelů.

b) Příprava datových kolekcí a pomocných rutin vyhledávacího systému ve vícejazyčných korpusech, včetně prostředků pro zpřesňování uživatelských dotazů pomocí thesauru a nástrojů pro disambiguaci víceznačných slov, na bázi klient/server aplikace. Jednotlivé dílčí výsledky řešení projektu lze charakterizovat takto:
- vytvoření multijazykových korpusů – základní výběr zahrnuje angličtinu a češtinu, dle možností alespoň některé úlohy plánujeme provádět i se slovenštinou (zajímavá je blízkostí k češtině) a němčinou,
- metoda automatického rozpoznání jazyka – kombinace „stop slov“ a frekvenčních znakových metod.

c) Příprava datových kolekcí a modulů pro filtraci a sumarizaci textů:
- vytvoření sumarizačních korpusů (pro angličtinu plánujeme využít standardních korpusů, např. DUC a pro češtinu bude vytvořen vlastní, složený vesměs z textů novinových článků,
- sumarizace textů založená na latentní sémantické analýze (LSA), vytvoření anotované kolekce pro sumarizátor založený na LSA
- vytvoření vícejazyčných korpusů,
- rozšíření standardních textových korpusů o korpusy závadných dokumentů pokrývající problematická témata definovaná v zadání.

d) Korpus syntaktických stromů (treebank):
- korpus bude morfologicky označkován a zjednoznačněn,
- bude v něm vyznačena závislostní struktura věty i jednotlivé větné složky včetně koreferenčními vztahy,
- korpus bude z části založen na existujícím PDT.

e) Korpus vzorových přepisů vybraných vět a jejich sémantické reprezentace:
– text korpusu bude podmnožinou korpusu syntaktických stromů,
– ve stromech budou vyznačeny významy z dostupných ontologií (WordNet),
– věty budou rozšířeny o logické formy.

f) Doplnění morfologického značkovače o robustní hádací proceduru, která bude spolehlivě přiřazovat morfologické značky i neznámým slovům.

3.3.3. Forma zpracování a předání výsledků dílčího cíle

Jedná se o vytvoření podpůrného aparátu, bez něhož nelze další zamýšlené cíle projektu dosáhnout. Vytvořeny budou proto korpusy v podobě rozsáhlých datových souborů se specifickou strukturou a organizací a pro jejich údržbu a prohledávání budou vyvinuty speciální softwarové nástroje. Výsledky budou soustředěny do soustavy datových souborů a její obsah prezentován formou publikace na konferencích a v průběžných výzkumných zprávách.

3.3.4. Kritické předpoklady dosažení dílčího cíle

Rizikové faktory ovlivňující náplň dílčího cíle „1“ a nástin jejich řešení jsou následující:

RF1: Během zpracování korpusů a korpusových nástrojů se vyskytnou další korpusy obsahující srovnatelná data.
Řešení: Korpusy pro český jazyk vznikají v ČR na celkem pěti pracovištích, která udržují těsné kontakty a výsledky výzkumu si vzájemně vyměňují nebo se o nich poměrně obsáhle informují. Navíc je třeba rozlišovat mezi korpusy psanými (textovými) a řečovými. Řečové korpusy vznikají prakticky jen na pracovištích v Plzni, Brně a Liberci, z nichž dvě se na řešení tohoto projektu budou podílet. Navíc vznik jakéhokoli dalšího korpusu je pozitivním jevem, neboť v tomto oboru více než kdekoli jinde platí, že vhodných dat není nikdy dostatek. Tudíž korpusy vytvořené v rámci navrhovaného projektu budou v každém případě využity i dalšími pracovišti. V případě cizojazyčných korpusů budou využívány korpusy, které jsou k dispozici v systému ELRA (European Language Resources Association).

RF2: Nepodaří se získat dostatek materiálů, resp. mluvčích, pro vytvoření textových, resp. audiokorpusů.
Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť již současný web poskytuje doslova nepřeberné množství textového materiálu, z nichž lze za použití vhodných vyhledávacích metod vybrat dostatečné množství materiálu pro vytvoření korpusu. V případě řečových korpusů nejde ani tak o problém nalezení vhodné množiny dat nebo množiny vhodných mluvčích, nýbrž kritickým místem je čas. Pořizování

řečových dat a zejména jejich následné zpracování (třídění, anotace, apod.) vyžaduje značné množství času, avšak riziko lze úspěšně odstranit kvalitním managementem projektu.

RF3: V průběhu naplňování dílčího cíle projektu se vyskytne komerční software řešící problematiku pořizování korpusů.
Řešení: Pokud se nějaký software vyskytne a bude využitelný, nebude díky modularitě předpokládaného programového vybavení příliš obtížné ho do vytvářeného software začlenit. Pravděpodobnost jeho výskytu v dohledné době je však minimální.

3.3.x. DÍLČÍ CÍL ŘEŠENÍ PROJEKTU - "2" - podrobně

3.3.1. Definice dílčího cíle

3.3.1.1. Co má být dílčím cílem dosaženo

Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka.

3.3.1.2. Datum dosažení dílčího cíle

31.12.2008

3.3.1.4. Typ činnosti při řešení dílčího cíle

Základní výzkum

3.3.2. Výsledky dílčího cíle

a) Návrh formalismu pro popis sémantiky na rozsáhlejší doméně, návrh vhodně strukturovaného sémantického popisu dotazů uživatelů, eventuálně vytvoření vlastního hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému.
b) Vytvoření ontologií pro aplikaci formalismu popisujícího sémantiku. Jednotlivými výsledky budou:
- návrh ontologie, sémantických konceptů – datový formát XML, vytvoření UML modelu,
- návrh ohodnocení jednotlivých konceptů vektorem sémantických příznaků, a to jak doménových, tak obecnějšího charakteru,
- návrh soustavy vektorů ohodnocení jednotlivých konceptů.
c) Vytvoření multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí, jeho zakomponování do prostředí pro vyhledávání a vývoj metod ohodnocování jeho kvality, návrh metod disambiguace v multijazykovém prostředí s využitím kontextu, thesauru a pravděpodobnostních metod:
- obohacený sumarizační systém o kompresi souvětí,
- systém rezoluce anafor a jeho využití při sumarizaci – pro angličtinu bude využit systém GuiTAR, vytvořený na univerzitě Essex (Anglie), pro češtinu bude na základě poznatků získaných na českých pracovištích vytvořen vlastní systém,
- metoda hodnocení kvality sumarizátorů na základě LSA.
d) Vývoj nových, dokonalejších modelů elektronických dokumentů tak, aby při použití textových klasifikačních algoritmů bylo dosaženo co nejlepších výsledků při rozpoznávání tématu, rozpoznávání spamových emailů, detekci dokumentů se závadným obsahem apod.
e) Vytvoření metodologie a nástrojů pro analýzu webových dokumentů.

3.3.3. Forma zpracování a předání výsledků dílčího cíle

Při naplňování tohoto dílčího cíle půjde o vytvoření základního teoretického podpůrného aparátu, bez něhož nebude možné další kroky realizovat. Jediný tento dílčí cíl bude mít charakter spíše základního výzkumu – půjde o vývoj metod, metodologií a formálních modelů pro návrh zamýšleného komunikačního rozhraní, avšak součástí výzkumných prací bude též experimentální implementace a vytvoření softwarových nástrojů pro evaluaci vyvíjených metod a formalismů. Výsledky budou shrnuty do písemných dokumentů a prezentovány téměř výhradně formou publikace na konferencích, v odborných časopisech a v průběžných výzkumných zprávách.

3.3.4. Kritické předpoklady dosažení dílčího cíle

Rizikové faktory ovlivňující dosažení dílčího cíle „2“ a nástin jejich řešení mohou být následující:

RF1: Nepotvrzení či neplatnost výzkumných hypotéz poskytujících základ pro vytvoření formalismů a modelů.
Řešení: Plánovaný dílčí cíl zde nestojí na jediné výzkumné hypotéze, nýbrž na teoretickém základu návrhu komunikačních systémů. Využito bude jak dosavadních poznatků z návrhu existujících komunikačních rozhraní a systémů pro interakci člověka s počítačem, tak i poznatků z psychologie komunikace a doporučení TC.13 IFIP (for HCI). Základním rizikem proto bude opět časový faktor, který lze výrazně omezit dobrým managementem projektu.

RF2: Nedostatečná erudice členů týmu pro vývoj formálních prostředků.
Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť oba participující týmy jsou složeny minimálně z poloviny ze starších zkušených výzkumníků, z nichž někteří se předmětnou oblastí zabývají 25 i více let, z druhé části pak z mladých perspektivních pracovníků, kteří buď vyrostli anebo se podíleli na řešení podobné problematiky a potřebné teoretické základy oboru již získali, zejména v doktorandském studiu.

3.3.x. DÍLČÍ CÍL ŘEŠENÍ PROJEKTU - "3" - podrobně

3.3.1. Definice dílčího cíle

3.3.1.1. Co má být dílčím cílem dosaženo

Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce.

3.3.1.2. Datum dosažení dílčího cíle

31.12.2009

3.3.1.4. Typ činnosti při řešení dílčího cíle

Aplikovaný výzkum s výjimkou průmyslového výzkumu (tzv. "neprůmyslový výzkum")

3.3.2. Výsledky dílčího cíle

a) Implementace uživatelského rozhraní pro hlasovou komunikaci se sémantickým webem –součástí výsledku budou:
- implementace LVCSR rozpoznávače,
- natrénování akustických a jazykových modelů,
- implementace nahrávacího modulu se stochastickým modelem detekce řečového signálu,
- implemetace parametrizátoru na bázi MFCC,
- návrh a implementace modulu pro akustické modelování založeného na umělých neuronových sítích nebo směsích Gaussových funkcí,
- návrh a implementace efektivního dekódovacího algoritmu, který dokáže pracovat s gramatikami a stochastickými jazykovými modely,
- programová realizace a ověření funkčních vlastností robustního ASR systému pro inflexní jazyky.

b) Systém pro extrakci významu ze spontánních promluv – dílčími kroky k dosažení tohoto dílčího cíle budou:
- návrh a realizace optimální řečové databáze,
- návrh systému sémantického značkování řečových dat,
- báze znalostí umožňující automatizované či automatické značkování spontánních promluv uložených v databázi,
- implementace stochastických sémantických gramatik pro automatickou sémantickou analýzu dotazu uživatele,
- využití hierarchické ontologie pro tvorbu strukturalizovaného popisu dotazů uživatele a pro zajištění schopnosti zobecňování z natrénovaných dat,
- aplikace metod mělkého (shallow) parsingu promluv pro částečnou analýzu dotazů uživatele.

c) Vytvoření komfortního uživatelského rozhraní pro práci se sémantickým webem – součástí tohoto dílčího cíle bude:
- návrh příslušného dialogového manageru akceptujícího tzv. kombinovanou iniciativu ve vedení dialogu (mixed initiative),
- vytvoření robustního systému pro efektivní a časově nenáročné vyhledávání dat v řečové databázi,
- vytvoření robustního a spolehlivého modelu sémantické hierarchie a jeho implementace.

d) Aplikace a modifikace OWL standardu v českém prostředí.

e) Aplikace klasifikačních metod v multijazykovém prostředí.

f) Kompletace multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí.

g) Algoritmy vhodné pro generování itemsetů a n-gramů a ověření jejich úspěšnosti pro klasifikaci textových dokumentů.

h) Výchozí algoritmy pro vyvozování nových znalostí z informací získaných z volného textu.

i) Prototyp programu pro přiřazování logických formulí větám z volného textu.

3.3.3. Forma zpracování a předání výsledků dílčího cíle

V dílčím cíli „3“ jde o vytvoření souboru programových produktů, které vzniknou implementací teoretických metod a formalismů vytvořených v rámci dílčího cíle „2“. Výsledky budou mít jednoznačně aplikační charakter, i když vesměs půjde jen o experimentální software, bez něhož nelze metody a modely verifikovat. Výsledky však budou moci být předávány i dalším zájemcům, protože se předpokládá úplná dokumentace vytvořeného programového vybavení. Výsledky budou proto prezentovány jednak jako balíky experimentálního software a metod, jednak budou také publikovány na konferencích, v průběžných výzkumných zprávách, případně také zveřejněny formou speciálních letáků, v tisku a uvažuje se též o možnosti předvedení na specializovaných veletrzích a výstavách.

3.3.4. Kritické předpoklady dosažení dílčího cíle

Rizikové faktory ovlivňující dosažení dílčího cíle „3“ a nástin jejich řešení:

RF1: V průběhu projektu přestane být o vytvářené přístupové technologie zájem a pracoviště účastnící se na řešení projektu se tak ocitnou bez reálné využitelnosti svých výsledků.
Řešení: Současným trendem je naopak příklon k využívání multimediálních a multimodálních dat, ukládání velkých množství dat a informací na běžných počítačových prostředcích, sílí propojování informačních technologií s rozhlasovým a televizním vysíláním, streamovanými médii a mobilními komunikacemi. Nové hardwarové prostředky budou vyžadovat nové technologie přístupu k datům, přičemž preferována bude komunikace v přirozeném jazyce, ať už psanou nebo mluvenou formou. Vyvíjené programové prostředky tento trend jednoznačně podpoří a proto je toto riziko za dobu řešení projektu téměř nulové.

RF2: V průběhu řešení projektu se vyskytne komerční software řešící problematiku srovnatelnou s

předpokládanými výsledky projektu.
Řešení: Komerční řešení využívající přístup k datům na webu prostřednictvím přirozeného jazyka jsou dosud v plenkách a komerční sféra naopak aktivně vyhledává zajímavé práce z akademické sféry. Proto je toto riziko minimální, očekáváme naopak velký zájem z komerční sféry.

RF3: Časové faktory ovlivňující zpracování software.
Řešení: Při implementaci a programové realizaci metod vyvinutých v rámci dílčího cíle „2“ může dojít k určité časové tísni vlivem nevhodně zvolených implementačních nástrojů, eventuálně nezkušeností některých mladších členů týmu. Riziko je však minimální, neboť řešitelský kolektiv je složen vesměs ze zkušených výzkumníků a mladých pracovníků, kteří již obdobné, i když jednodušší systémy v minulosti vytvářeli a implementovali. Časový faktor lze navíc výrazně ovlivnit dobrým managementem projektu.

3.3.x. DÍLČÍ CÍL ŘEŠENÍ PROJEKTU - "4" - podrobně

3.3.1. Definice dílčího cíle

3.3.1.1. Co má být dílčím cílem dosaženo

Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí.

3.3.1.2. Datum dosažení dílčího cíle

31.12.2010

3.3.1.4. Typ činnosti při řešení dílčího cíle

Aplikovaný výzkum s výjimkou průmyslového výzkumu (tzv. "neprůmyslový výzkum")

3.3.2. Výsledky dílčího cíle

a) Testování a ověřovací provoz implementovaného hlasového rozhraní – součástí bude
- otestování zpracovaného LVCSR rozpoznávacího systému,
- ověření funkčních vlastností robustního ASR systému na vhodné množině uživatelů,
- otestování vyvinutých metod automatické sémantické analýzy dotazů.

b) Ověření funkčních vlastností vytvořených ontologií a hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému analýzy sémantiky,

c) Ověření vlastností algoritmů pro klasifikaci a analýzu dat na různých typech dokumentů.

d) Otestování a ověření navržených metod na konkrétních typových řešeních, např. na přístupu k webovým stránkám výzkumných a vzdělávacích institucí.

e) Na základě poznatků získaných z předchozích bodů jasně definované nejvhodnější přístupy k řešení zvolených problémů, doložené výsledky z aplikace jednotlivých postupů do praxe.

f) Vyhodnocení úspěšnosti jednotlivých fází analýzy volného textu od morfologické úrovně až po převod do logických formulí.

3.3.3. Forma zpracování a předání výsledků dílčího cíle

Náplní dílčího cíle „3“ je provedení rozsáhlých testů (tzv. field experiments) vyvinutých metod, metodologií, modelů a vytvořeného souboru programových produktů. Předpokládá se testování produktů na obvyklých třech skupinách uživatelů – v prvním kroku budou vlastnosti systémů a metod prověřovány úzkou skupinkou řešitelů projektu, ve druhém kroku bude testovací množina uživatelů vytvořena ze spolupracovníků, kteří však s řešením projektu neměli nic společného a o výsledcích řešení jsou jen velmi kuse informováni, a teprve ve třetím kroku bude systém testován libovolnými uživateli, tzv. „lidmi z ulice“. Zčásti však v tomto kroku budou využiti studenti, kteří všeobecně mají tendenci takové systémy „pokořit“. Výsledky budou kompletně dokumentovány a z vyhodnocení experimentů budou vyvozovány příslušné závěry, tj. systém a jeho části budou průběžně doplňovány, upravovány a opětovně testovány. V závěru budou výsledky testování a ověřovacího provozu publikovány v časopisech, na konferencích a obšírně v závěrečné výzkumné zprávě.

3.3.4. Kritické předpoklady dosažení dílčího cíle

Rizikové faktory ovlivňující dosažení dílčího cíle „4“ a možná řešení:

RF1: V průběhu testů se projeví nedostatky v koncepci systému vedoucí k závažným problémům ve funkci systému.
Řešení: Řešitelský tým je složen z odborníků, kteří obdobné, i když jednodušší, systémy již vytvořili a mají z jejich tvorby nezanedbatelné zkušenosti. Tým byl dále doplněn o mladé pracovníky, kteří se podíleli na tvorbě řady produktů pro prezentace na webových stránkách a je jim problematika přístupu k webu velmi blízká. Riziko volby nevhodné koncepce je proto minimální.

RF2: V průběhu testů se projeví nedostatky v implementaci systému a metod.
Řešení: Obdobné jako předchozí rizikový faktor – řešitelský tým je složen z odborníků, kteří obdobné, systémy již vytvořili a mají i z jejich implementace poměrně rozsáhlé zkušenosti. Riziko závažných implementačních chyb je proto minimální, drobné nedostatky v implementaci bývají zpravidla v krátké době snadno odstranitelné.

RF3: Nepodaří se vytvořit dostatečně reprezentativní množiny testovacích osob.
Řešení: Ve vztahu k odstavci 3.3.3. (tři úrovně testování) je riziko nedostatečného vytvoření skupin testujících osob nepatrné – obě participující pracoviště jsou poměrně rozsáhlá a množinu osob testujících vlastnosti systému nebude problém vytvořit; ostatně bylo již ověřeno v minulosti na jednodušších úlohách. Otázka volby třetí skupiny osob je spíše otázkou vytvořeného přístupu k systému – zde se nabízejí dvě možnosti: Buď si osoby vhodné k testování systému vybírat podle určitých hledisek (bylo tak někdy postupováno v minulosti a osoby byly k testování zvány na řešitelské pracoviště) nebo zveřejnit přístupový portál systému a dovolit testování systému široké veřejnosti prostřednictvím internetu, popř. přes telefon (telefonní přístup je však v současných podmínkách omezen kvalitou spojení v mobilních sítích, resp. kvalita spojení je dána úrovní signálu v místech, kde se potenciální uživatel právě nachází, a výsledky testů jím mohou být zkresleny). Rizikový faktor může být opět minimalizován vhodnými rozhodnutími, resp. dobrým managementem projektu.

4. PLÁN PROJEKTU

4.1. Metodika řešení

Zatímco v odstavci 3.3 byly popsány parciální cíle členěné z věcného a časového hlediska, obsahuje tento odstavec popis postupů, jak bude dílčích cílů dosaženo a jakým způsobem budou výsledky každého z nich po ověření v závěrečné testovací fázi začleněny do modulů finálního souboru nových lingvistických prostředků pro komunikaci se sémantickým webem.
Uživatelské rozhraní pro hlasový vstup/výstup, které bude použito pro komunikaci se sémantickým webem, a související robustní ASR systém pro inflexní jazyky budou vytvořeny na základě kvalitního korpusu pro ASR a z něj extrahovaného dostatečného množství trénovacích dat. Korpus se bude členit na dvě části uvedené v dílčím cíli „1“ a poslouží též pro natrénování akustických modelů rozpoznávače.
Pro nalezení formalismu pro popis sémantiky na rozsáhlejší doméně bude třeba navrhnout vhodně strukturovaný sémantický popis dotazů uživatelů, ověřit možnost použití existujících ontologií či vytvoření vlastního hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému, provést anotaci testovacího korpusu zvoleným sémantickým popisem, implementovat a otestovat několik metod automatické sémantické analýzy dotazů a experimentálně implementovat systém schopný tento popis strojově generovat. Pokusíme se proto o vytvoření ontologií pro aplikaci formalismu popisujícího sémantiku. K tomu bude využito sémantických konceptů – datový formát XML, vytvoření UML modelu, návrhu systému ohodnocení jednotlivých konceptů vektorem sémantických příznaků, a to jak doménových, tak obecnějšího charakteru, vytvoření speciálního systému značkování řečové databáze a návrhu soustavy vektorů ohodnocení jednotlivých konceptů na vstupu samoorganizující se neuronové sítě (Kohonenova mapa + případná modifikace) – budou provedeny experimenty se vstupními vektory a samotnou organizací neuronové sítě. Uvažuje se rovněž o návrhu soustavy vektorů ohodnocení jednotlivých konceptů na vstupu neuronové sítě využívajících jako základní výpočetní jednotku synapsi.
K vytvoření systému pro extrakci významu ze spontánních promluv bude navržena, implementována a realizována (postupně naplněna daty) řečová databáze a související systém vyhledávání v ní, navržen systém sémantického značkování řečových dat a automatizované či automatické značkování spontánních promluv uložených v databázi bude prováděno prostřednictvím báze znalostí, což bude jedním z významných výsledků projektu.
Vytvoření komfortního uživatelského rozhraní pro práci se sémantickým webem bude vyžadovat kromě návrhu příslušného dialogového manageru akceptujícího tzv. kombinovanou iniciativu ve vedení dialogu (mixed initiative) vytvoření robustního systému pro efektivní a časově nenáročné vyhledávání dat v řečové databázi a dále návrh a implementace robustního a spolehlivého modelu sémantické hierarchie.
Vytvoření vyhledávacího systému ve vícejazyčných korpusech, včetně prostředků pro zpřesňování uživatelských dotazů pomocí thesauru a nástrojů pro disambiguaci víceznačných slov, busde realizováno na bázi klient/server aplikace. Toto bude řešeno na základě přípravy multijazykových korpusů, zpřesnění metod automatického rozpoznání jazyka, vývoje speciálních metod disambiguace v multijazykovém prostředí s využitím kontextu, thesauru a pravděpodobnostních metod, implementace klasifikačních metod v multijazykovém prostředí a aplikace specifických metod vyhledávání v multijazykovém prostředí. S využitím existujících vyhledávačů se budeme snažit modifikovat dotaz prostřednictvím thesauru do dalších jazyků se současným řešením problému zjednoznačnění významu termů.
Vytvoření multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí, jeho zakomponování do prostředí pro vyhledávání a vývoj metod ohodnocování jeho kvality bude vyžadovat přípravu sumarizačních korpusů, přičemž pro angličtinu plánujeme využití standardních korpusů (např. DUC) a pro češtinu vytvoříme vlastní. Sumarizace založená na latentní sémantické analýze (LSA), která je schopna zachytit témata textu, bude implementována pro její jazykovou nezávislost. Na základě LSA bude rovněž vyvinuta metoda hodnocení kvality sumarizátorů.
Vývoj nových, dokonalejších modelů elektronických dokumentů a jejich implementace tak, aby při použití textových klasifikačních algoritmů bylo dosaženo co nejlepších výsledků při rozpoznávání tématu, rozpoznávání spamových emailů, detekci dokumentů se závadným obsahem apod. bude uskutečněn na bázi rozšíření standardních textových korpusů (potřebných pro průběžné porovnávání dosažených výsledků) o korpusy závadných dokumentů pokrývající problematická témata definovaná v zadání. Tyto korpusy jsou potřebné pro trénování a testování metod a v současné době chybí. Dále budou navrženy vhodné algoritmy pro generování itemsetů a n-gramů a prověřen jejich vliv na úspěšnost klasifikace textových dokumentů pro různé přístupy k výběru kandidátních položek. Bude zapotřebí vzájemně porovnat vliv itemsetů a n-gramů na úspěšnost klasifikace, pokusit se zkombinovat oba přístupy do modelu dosahujícího lepších výsledků. Zkoumány budou proto různé textové klasifikátory a experimentálně bude ověřována jejich úspěšnost při klasifikaci textových dokumentů, webových dokumentů a spamových e-mailů vzhledem k jejich reprezentaci navržené výše.
Vytvořená metodologie a nástroje pro analýzu webových dokumentů budou využity k evaluaci institucí (vzorovým případem bude hodnocení kateder/ústavů ve VaV), příp. i jednotlivců. Postupovat se bude podle metody, jež ze zadaného uzlu ve webovém grafu na základě jeho okolí najde komunitu relevantních webových stránek. Parametry budou velikost komunity, hloubka prohledávání apod. Metodami strojového učení extrahujeme relevantní informace (jména, názvy, místní názvy apod.) ze stažených dokumentů, navrhneme a

implementujeme algoritmy kombinující citační analýzu webových stránek s informacemi extrahovanými z jejich obsahu za účelem evaluace subjektů reprezentovaných danými stránkami.
Jednotlivé programové moduly předpokládaného programového systému budou důsledně zpracovávány technologií tvorby inteligentních agentů s cílem dosažení maximálního stupně modularity a otevřenosti programového systému. Zvolená technologie programování rovněž umožní použití vytvořených programových modulů v dalších aplikacích, resp. umožní vytváření aplikací "šitých na míru" podle konkrétních požadavků uživatele / zadavatele.
Metody pro strukturní analýzu volných textů, zahrnující syntaktickou analýzu, logickou analýzu věty pro reprezentaci znalostí a vyvozování nových faktů na základě automaticky tvořené báze těchto znalostí, vyžadují v první fázi vytvoření několika strukturních jazykových zdrojů. Konkrétně se jedná o korpus syntaktických stromů, tzv. treebank a korpus vzorových přepisů vybraných vět na jejich sémantické reprezentace. Při budování těchto zdrojů bude použito maximum zdrojů a postupů získaných při práci s podobnými materiály dříve (např. textové korpusy na FI MU nebo pražský stromový korpus PDT).
Získané podkladové jazykové zdroje budou využity jako trénovací a testovací data pro konkrétní metody syntaktické a logické analýzy. Nejdůležitějším prvkem těchto analýz bude postupný vývoj rozsáhlé strojové gramatiky české syntaxe, předpokládá se použití mechanizmu metagramatiky využívající kontextových podmínek a akcí. Důležitou součástí analýzy bude zapojení informací z dalších stávajících jazykových zdrojů jako je morfologický analyzátor, sémantická databáze WordNet nebo lexikon slovesných valencí.
Logická analýza věty představuje v současnosti možnost nejúplnější formy pro reprezentaci významu textových i multimodálních dat. Nejrozšířenější reprezentací znalostí v oblasti sémantického webu je v současnosti OWL+RDF, kde vyjadřovací síla tohoto formalismu odpovídá predikátové logice prvního řádu. Pro zachycení úplného spektra fenoménů přirozeného jazyka se jeví jako vhodné použít formalismus, který bude zdola kompatibilní s OWL, ovšem bude založen na logické teorii vyšších řádů umožňující popsat temporalitu a intenzionalitu analyzovaných pojmů (např. Montagueho intenzionální logika nebo Transparentní intenzionální logika). Algoritmus pro vyvozování nových znalostí lze díky uvedené zpětné kompatibilitě budovat postupně, s tím, že bude nejprve ověřen a implementován algoritmus založený na rezoluci v predikátové logice. Tento algoritmus bude pak rozšířen na práci s kompletním arzenálem použitého logického formalismu, vhodným základem se v současnosti jeví Gentzenův kalkul sekventů.
V průběhu práce bude správnost a kvalita použitých formalismů, metod a algoritmů opakovaně prověřována na testovacích datech a získané parametry budou srovnávány s jinými projekty podobného zaměření v ČR i v zahraničí. Důležitým kritériem úspěšnosti těchto metod bude i jejich vzájemné propojení, kdy výsledky syntaktické analýzy slouží jako základ pro reprezentaci znalostí a ta pak je klíčová pro vyvozování nových faktů.
Vstupní data pro vývoj a analýzu dialogových rozhraní založených na sémantickém webu pro aplikace v asistivních technologiích budou představovat zejména dialogové korpusy vytvořené simulováním reálných dialogů metodou Wizard of Oz. Vývoj metod pro konverzi grafických uživatelských rozhraní do dialogové podoby s využitím prostředků sémantického webu bude podporován rozsáhlou grafickou objektovou databází vyvíjenou ve spolupráci s Laboratoří vyhledávání dat Fakulty informatiky MU Brno.
Při analýze a implementaci dialogových systémů budou využívány matematické modely založené na konečněstavové analýze dialogů. Informační dostupnost (accessibility) a metody vytváření webových stránek pomocí dialogu pro nevidomé budou vyhodnocovány a testovány ve spolupráci se střediskem MU pro pomoc studentům se specifickými nároky.

4.2. Projektový tým a řešitelské týmy

4.2.1. Představení projektového týmu

Projektový tým je sestaven z pracovníků Katedry informatiky a výpočetní techniky Fakulty aplikovaných věd Západočeské univerzity v Plzni a Centra zpracování přirozeného jazyka Fakulty informatiky Masarykovy univerzity v Brně. Při sestavování týmu byla vzata v úvahu jak výborná úroveň vztahů mezi oběma pracovišti, tak i potřeba složit tým jak z odborníků na informatickou vědu s inklinací k využívání metod umělé inteligence, tak z odborníků na počítačovou lingvistiku.
Výzkumná skupina na KIV FAV ZČU se dlouhodobě (od r. 1981) věnuje tématice vztahující se k výzkumnému zaměření navrhovaného projektu. Řešitelský kolektiv je tvořen zčásti habilitovanými pracovníky a graduovanými odbornými asistenty (PhD.), zčásti pak doktorandy garantů projektu. První výzkumná skupina se od r. 1993 zabývá problematikou vývoje dialogových informačních systémů a v posledních létech úspěšně vytvořila několik funkčních aplikací. V rámci projektu se bude orientovat na vývoj uživatelsky přívětivého komunikačního rozhraní pro komunikaci s webem a také analýzou dat reprezentovaných v přirozeném jazyce. Druhá skupina řeší problematiku zpracování a využívání textových informací pro různé účely. Do oblasti jejího zájmu patří zejména metody vyhledávání, sumarizace a filtrace dokumentů a jejich využití jak v digitálních knihovnách, tak v prostředí www. Funkční kategorizační systém pracující na bázi metody navržené tímto kolektivem je užíván v oborové knihovně ZČE. V této skupině se počítá i s účastí 4 mladých perspektivních pracovníků, kteří budou v letošním roce předkládat doktorské disertace.

Řešitelský kolektiv FI MU Brno bude tvořen pracovníky Centra zpracování přirozeného jazyka. Toto centrum organizačně zahrnuje dvě výzkumné laboratoře – laboratoř zpracování přirozeného jazyka (LZPJ) a laboratoř řeči a dialogu (LSD). Laboratoř zpracování přirozeného jazyka se dlouhodobě podílí na výzkumu a experimentálním vývoji ve všech základních oblastech počítačové lingvistiky. V LZPJ byl vytvořen vícejazyčný morfologický analyzátor AJKA, který pro zadané slovo určí vzor, základní tvar, všechny gramatické kategorie, segmentaci a případné slovotvorné vazby na jiná slova. Analyzátor spolupracuje s editorem morfologické databáze I_PAR, který umožňuje zadávání vztahů mezi tvarotvornými a slovotvornými vzory a vytváření slovotvorných vzorů. LZPJ je autorem české lexikální databáze typu WordNet i obecného nástroje na editaci lexikálních databází uložených ve formátu XML (VisDic) používaného v několika zemích EU. Na poli syntaktické analýzy češtiny v LZPJ vzniká analyzátor SYNT, který v současnosti dosahuje přes 90% pokrytí na volných textech. V návaznosti na SYNT se pracuje na algoritmu vytvářejícím v omezeném rozsahu sémantické reprezentace českých vět na bázi transparentní intenzionální logiky (NTA - normální translační algoritmus). V oblasti korpusů a korpusových nástrojů se Laboratoř zpracování přirozeného jazyka FI MU (LZPJ) komplementárně podílí na budování Českého národního korpusu. Na základě výzkumu v LZPJ vznikl korpusový manažer Manatee, který umožňuje zpracovávat rozsáhlé korpusy s velikostí přes miliardu slovních tvarů. Systém je nyní v rutinním provozu na několika pracovištích v ČR i v zahraničí. Laboratoř řeči a dialogu se dlouhodobě zabývá výzkumem v oblasti dialogových systémů, syntézy a rozpoznávání řeči a dále také vývojem metod pro indexaci a vyhledávání v multimediálních datech. Jedním z přínosů činnosti pracoviště je výzkum v oblasti optimalizace dialogových strategií, jeho výsledky lze použít i v oblasti multimodální komunikace. K rozšíření dialogových systémů o různé formy komunikace lze použít některé výsledky dosažené především v laboratořích zpracování přirozeného jazyka, interakce člověka s počítačem a laboratoře pokročilých síťových technologií. Jedním z hlavních výsledků výzkumu za poslední období je vytvoření dialogové platformy OptimTalk pro interpretaci jazyka VoiceXML. V rámci laboratoře byl rovněž vyvinut syntetizér řeči Demosthenes.

Celkem se počítá s účastí 22 výzkumníků na řešení projektu.

4.2.2. Řešitelská pracoviště

Seznam jednotlivých pracovišť

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

Masarykova univerzita
Fakulta informatiky
Centrum zpracování přirozeného jazyka

Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

4.2.2.1. - Údaje o pracovišti - Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

Název pracoviště

Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

Popis pracoviště podílejícího se na činnosti centra , jeho lokalizace a popis prostor vymezených pro jeho činnost

Pracoviště je umístěno v budově Fakulty aplikovaných věd Západočeské univerzity v Plzni, čtvrté podlaží. Sestává z objektu kanceláří pracovníků, objektu laboratoří a výukového objektu. Projekt bude v případě jeho přijetí řešen zčásti v kancelářích pracovníků, zčásti pak v laboratorním objektu, konkrétně v Laboratoři inteligentních komunikačních technologií (místnosti UL 410 a UL 412). Garant projektu doc. Ježek je dislokován v kanceláři UK 428, garant prof. Matoušek pak v kanceláři UK 424. Výzkumná činnost bude probíhat kromě zmíněných laboratorních prostorů částečně též v pracovnách odborných asistentů a doktorandů (UU 404, UK 415). Materiálně je pracoviště plně zajištěno výpočetní technikou (osobní počítače, notebooky, pracovní stanice a výkonný serverový cluster s diskovými poli o kapacitě 3,5 TB) a běžnou kancelářskou technikou jako jsou tiskárny, kopírky, telefax, atd.

Dosavadní výzkumná činnost

Dosavadní podíl pracoviště na řešení VV úkolů v národním a mezinárodním kontextu můžeme dokumentovat následujícími skutečnostmi. V uplynulých letech se pracoviště účastnilo řešení 9 projektů na národní úrovni, včetně výzkumného záměru, a pěti mezinárodních projektů, včetně dvou projektů Evropské unie. Výsledky výzkumné práce v uplynulých pěti letech byly publikovány ve 264 článcích v časopisech a sbornících konferencí na národní a mezinárodní úrovni. Své práce obhájilo 17 studentů doktorského studia. Řešitel má úzkou spolupráci s univerzitami a výzkumnými institucemi na národní i mezinárodní úrovni. Řada z dosažených výsledků výzkumu našla využití v aplikacích. Pracovníci řešitele jsou členy redakčních rad vědeckých časopisů i programových výborů národních i mezinárodních konferencí.
Vyvinut byl Experimentální Deduktivní Databázový systém (EDD), založený na zobrazení rozšířeného logického databázového jazyka Datalog do SQL. Rozšíření zahrnují imperativní konstrukce, agregáty, negaci a neurčitost faktů i pravidel, která dovolují kombinovat deduktivní vlastnosti spolu s relačními a zpracovávat rozsáhlá množství dat s použitím znalostních principů. Původním výsledkem je i formální důkaz úplnosti a správnosti konstrukce pravidel v optimalizační metodě magických množin při zahrnutí neurčitosti do procesu dedukce. EDD systém byl použit k vyšetřování závislostí v databázi studijní agendy. Na principu induktivního strojového učení byla vyvinuta a implementována původní metoda pro kategorizaci textových dokumentů. “Metoda Itemsets” je založena na detekci množin položek (termů) a rozšiřuje tradiční aplikace “Algoritmu apriori”. NBCI metodu lze charakterizovat jako metodu Itemsets využívající Naivní Bayesův klasifikátor. Obě metody jsou robustní, efektivní, navržené pro klasifikaci zejména krátkých dokumentů (abstraktů, souhrnů) a jejich principy jsou využity v informačním systému firmy InSITE.
V oblasti vývoje dialogových informačních systémů spolupracoval tým katedry informatiky a výpočetní techniky s pracovišti (Univerzita Erlangen, TU Košice, Univerzita Ljubljana) v rámci Evropského projektu z programu Copernicus (JRP No. 1634 ”SQEL”) a vývoji multilinguálního dialogového informačního systému poskytujícího informace o příjezdech a odjezdech vlaků, resp. příletech a odletech letadel po telefonu, a to ve čtyřech, resp. pěti evropských jazycích. Dvoujazyčný (česky a německy hovořící) městský, resp. obecní, informační systém poskytující turistické a ekonomické informace a dále informace o lokalizaci úřadů, úředních hodinách institucí, otevírací době velkých obchodů, lokalizaci a otevírací době muzeí či programech kin a divadel byl v poslední době vyvinut ve spolupráci katedry s univerzitami v Drážďanech a Řezně pro partnerská města Plzeň a Regensburg. V rámci projektu byl vyvinut originální robustní rozpoznávač fonetických segmentů, jehož využitím při implementaci informačního systému bylo dosaženo mimořádně vysoké spolehlivosti rozpoznávání uživatelových promluv. Katedra se rovněž účastnila řešení či koordinovala projekty sponzorované Grantovou agenturou České republiky (v posledních létech projekty GA 201/99/1248 “Dialogový systém pro programování zrakově postižených“ a GA 201/02/1553 “Komunikace s informačními databázovými systémy pro zdravotně a tělesně postižené na bázi přirozeného jazyka“), MŠMT v rámci programu KONTAKT a Fondem rozvoje vysokých škol.

Popis výzkumné činnosti pracoviště a jeho vybaveni

Katedra informatiky a výpočetní techniky dosáhla v oblasti vztahující se k cílům projektu těchto významných výsledků:
1. Experimentální Deduktivní Databázový systém (EDD), založený na zobrazení rozšířeného logického databázového jazyka Datalog do SQL. EDD dovoluje predikátovou formu popisu dotazů, včetně rekurzivních vazeb, imperativních predikátů, agregačních formulí a zavedení neurčitosti v extenzionální i v intenzionální části. Zpracování dotazu je možné optimalizovat „metodou magických množin“ a teoretické zdůvodnění optimalizačního postupu při zpracování neurčitosti je významným původním výsledkem práce. Citace: Ježek, K., Zíma, M.: Magic Sets Method with Fuzzy Logic, ADVIS 2002, Lecture Notes in Comp.Sc.2457 pp.83-92, Springer-Verlag 2002
2. Metoda „Itemsets” je novou kategorizační metodou vhodnou pro zpracování krátkých dokumentů. Byla prověřena její použitelnost i pro sumarizační účely, pro filtraci nevyžádané pošty a její kombinace s Bayesovským klasifikátorem. Nasazena fi. InSITE v oborové digitální knihovně. Citace: Hynek, J., .Ježek, K.: Use of Text Mining Methods in a Digital Library, Proc. ELPUB 2002, pp.276-286, Verlag fur Wissenschaft und Forschung Berlin 2002
3. Multilinguální multifunkční dialogový informační systém umožňující telefonický přístup k letovým a vlakovým jízdním řádům - systém byl vyvinut v rámci programu Copernicus (SQEL) a poskytuje uživateli informace o vlakových spojeních a letových řádech v němčině, angličtině, češtině, slovenštině a slovinštině na základě klasické telefonické komunikace v přirozeném jazyce. Unikátním výsledkem je modul adaptace systému na jazyk volajícího a uživatelsky přívětivé řízení

dialogu. Citace: Aretoulaki, M., Gallwitz, F.: Harbeck, S., Ipšič, I., Ivanecký, J., Matoušek, V., Niemann, H., Nöth, E., Pavešič, N.: SQEL - A Multilingual and Multifunctional Dialogue System. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP ´98), Sydney, Australia, volume 7, pp. 2283 - 2286, December 1998.
4. Piezoelektrické pero pro rozpoznávání ručně psaného textu (Smart P3) - systém pro autentikaci a autorizaci jedinců na základě verifikace pravosti podpisů pořízených speciálním biometrickým perem. Hardwarové řešení bylo realizováno na univerzitě v Regensburgu, na českém pracovišti byly vyvinuty metody extrakce příznaků z biometrických signálů a klasifikátor založený na využití neuronové sítě ART-2. Navržený systém má bezprostřední využití v bankovnictví, administrativní sféře, medicíně apod. Citace: Mautner P., Rohlík O., Matoušek V., Kempf J.: Signature Verification Using an ART-2 Neural Network. In: Proceedings of the Int. Conf. ICONIP ´02, Singapore, November 2002, pp. 374 – 381.
5. Robustní klasifikátor slov pro rozpoznávání spontánních promluv – byl vyvinut pro účely rozpoznávání spontánních promluv při komunikaci uživatele s obecním/městským informačním systémem poskytujícím informace z různých domén. Klasifikátor je koncipován jako hybridní struktura založená na kombinaci klasických statistických metod (HMM) a umělé neuronové sítě typu vícevrstvý perceptron. Využit byl při návrhu dvoujazyčného městského informačního systému pro partnerská města Plzeň a Regensburg. Citace: Ekštein, K.; Matoušek, V.; Mouček, R.: Hybrid analytic/ANN-based acoustic-phonetic decoding. In: Elektronische Sprachsignalverarbeitung, Tagungsband der 14. Konferenz, Dresden, w.e.b. Universitätsverlag & Buchhandel, 2003, S.75-81

Prostorové požadavky projektu na pracovišti řešitele jsou zabezpečeny v plném rozsahu. Řešitel má k dispozici také základní technické vybavení, které bude v rámci řešení projektu nutno doplnit a modernizovat.
Technické vybavení výzkumných laboratoří řešitele je následující:

- pracovní stanice (SUN, HP, Dell, Intergraph)
- PC: až 2,8 GHz, až 4GB RAM, několik se dvěma procesory, Windows XP a Linux operační sytémy
- 2 systémy na 64b platformách
- několik serverů se dvěma a čtyřmi procesory, až 4GB RAM
- zařízení pro virtuální realitu SuperScape, LCD Crystal Eye, datové rukavice, sledovač pohybu
- velkoplošná stereoskopická projekce

Řešitel sdílí specializovaný software v rámci univerzitní počítačové sítě

- OpenView Network Node Manager pro správu a monitorování počítačových sítí
- Xilinx Foundation 4.1 software
- Mentor Graphic FPGA Advantage
- Mentor Graphics Expedition PCB
- vývojové systémy pro analogová zařízení SHARC DSP
- systém pro vývoj aplikací virtuální reality SuperScape VRT
- modulární prostředí pro vizualizaci vyvinuté skupinou počítačové grafiky řešitele
- databázové systémy Oracle, Interbase, PostgreSQL
- vývojové nástroje Delphi, Visual C++, JBuilder, ...
- Nag Explorer & Modular Visualization Environment
- Lab Windows ICVI programming Environment
- Together6ControlCenter (CASE nástroj) a PowerDesigner7

Katedrová 100Mbps síť je propojena s univerzitní sítí, CESNETem a Internetem.

Katedra informatiky a výpočetní techniky využívá distribuované výpočetní prostředí zahrnuté do projektu Orion zhruba od roku 1999. Toto prostředí je založeno na bázi projektu MIT “Athena” a poskytuje uživatelům unifikovaný přístup ke všem výpočetním zdrojům a síťovým službám, včetně unifikovaného uživatelského rozhraní, autentikace přístupu a aplikačních databází. Uživatelé mají možnost přístupu do jednotného prostředí a využívání zdrojů buď ze standardních osobních počítačů nebo prostřednictvím výkonných pracovních stanic, nezávisle na jejich vlastní architektuře. Podporována jsou softwarová prostředí UNIX a MS Windows NT nebo XP, pro získání přístupu ke všem síťovým i výpočetním zdrojům stačí jediné přihlášení (logon) na libovolný zdroj distribuovaného prostředí.
Jako speciální subsystém projektu Orion byl v roce 2002 vytvořen cluster pracovních stanic Sun Blade 100 podporovaný dvěma servery – Sun Enterprise 250 a Sun Fire 280R. Oba servery jsou plnohodnotně začleněny do katedrální počítačové sítě a tvoří klesický serverový cluster se všemi možnostmi zástupné funkce či zálohování. Instalovaná disková kapacita obou serverů je momentálně tvořena dvěma diskovými poli o celkové kapacitě 3,5 TB.

Vyjádření k duplicitě řešení

Katedra informatiky a výpočetní techniky FAV ZČU v Plzni není v současné době zapojena do řešení významnějších projektů, center či výzkumných záměrů; v současné době je zpracováván a bude podán návrh výzkumného záměru s názvem „Perspektivní informační technologie“, jehož součástí bude vývoj obecné reprezentace sémantiky vět přirozeného jazyka. Jelikož se však bude jednat o čistě základní výzkum, jehož pravděpodobné výsledky budou v tomto navrhovaném projektu dále rozpracovány a aplikovány, nebude činnost v rámci navrhovaného projektu nijak duplicitní, naopak oba projekty se budou vhodně doplňovat, resp. na sebe vhodně navazovat.

4.2.2.2. Řešitelský tým - Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

Ježek Karel doc. Ing. CSc. 420617/110 Česká republika
řešitel koordinátor
Andrš David Ing. 800204/2653 Česká republika
člen řešitelského týmu
Beneš Vilém Ing. 800318/2220 Česká republika
člen řešitelského týmu
Ekštein Kamil Ing. PhD. 770530/2011 Česká republika
člen řešitelského týmu
Fiala Dalibor Ing. 800323/5845 Česká republika
člen řešitelského týmu
Klečková Jana doc. Dr. Ing. 496108/095 Česká republika
člen řešitelského týmu
Konopík Miloslav Ing. 810326/1782 Česká republika
člen řešitelského týmu
Krutišová Jana Ing. 595516/0046 Česká republika
člen řešitelského týmu
Matoušek Václav prof. Ing. CSc. 480613/108 Česká republika
člen řešitelského týmu
Mautner Pavel Ing. PhD. 650522/2592 Česká republika
člen řešitelského týmu
Mouček Roman Ing. PhD. 760707/2000 Česká republika
člen řešitelského týmu
Pavelka Tomáš Ing. 790918/2083 Česká republika
člen řešitelského týmu
Steinberger Josef Ing. 790918/2127 Česká republika
člen řešitelského týmu
Tesař Roman Ing. 790930/2379 Česká republika
člen řešitelského týmu
Toman Michal Ing. 800704/2054 Česká republika
člen řešitelského týmu

Ježek Karel doc. Ing. CSc. 420617/110 Česká republika
řešitel koordinátor
docent, vedoucí katedry
377 632 475 jezek_ka@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

Plánování a koordinace výzkumných prací, vedení kolektivu na KIV, pravidelná kontrola a hodnocení výsledků, vedení a výchova doktorandů

Prokázání odborné způsobilosti (seznam publikací)

1 - Hynek, J., Jezek, K., Rohlík, O.: Short Document Categorization - Itemsets Method, PKDD 4-th European Conference on Principles and Practice of Knowledge Discovery in Databases, Workshop Machine Learning and Textual Information Access. Lyon,France, Sept.2000, pp.14-19
2 - Ježek, K., Zíma, M.: Magic Sets Method with Fuzzy Logic, ADVIS 2002, Lecture Notes in Comp.Sc.2457 pp.83-92, Springer-Verlag 2002
3 - Steinberger, J., Jezek, K.: Text Summarization and Singular Value Decomposition, ADVIS 2004, Lecture Notes in Comp.Sc.2457 pp.245-254, Springer-Verlag 2004,
4 - Tesar, R., Fiala, D., Rousselot, F., Jezek, K.: A comparison of two algorithms for discovering repeated word sequences, Data Mining IV 2005, pp. 121-131, Skiathos, Greece, WIT Press, WIT Transaction on Information and Communication Technologies
5 - Jezek, K., Toman, M.: Documents Categorization in Multilingual Environment, ElPub2005, pp.97-104, Leuven, Belgium 2005, Peeters Publishing

Andrš David Ing. 800204/2653 Česká republika
člen řešitelského týmu
výzkumný pracovník
377 632 491 andrsd@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj a implementace analytických a klasifikačních metod, analýza a rekonstrukce rozpoznávaných větných struktur

Prokázání odborné způsobilosti (seznam publikací)

1 - Andrš, D.; Mouček, R.: Word extraction based on subword models. In: Proc. of 5th International PhD Workshop on Systems and Control, Hungary, September 2004, pp. 1 – 4.
2 - Mouček, R.; Andrš, D.: Theory of microsituations in spoken language. In: Proc. of 5th International PhD Workshop on Systems and Control, Hungary, September 2004, pp. 83 – 86.
3 - Andrš, D.: Using principles of Language Modeling in Decoding. In: Proc. of the 6th International PhD. Workshop on Systems and Control, Slovenia, October 2005.
4 - Pavelka, T.; Ekštein, K.; Andrš, D.: Hybridní rozpoznávač přirozené řeči pro český jazyk, Kognícia a umelý život V, Slovensko, květen 2005.

Beneš Vilém Ing. 800318/2220 Česká republika
člen řešitelského týmu
výzkumný pracovník
377 632 491 shodan@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj a implementace metod vyhledávání a korekcí dat vybavovaných z řečové databáze

Prokázání odborné způsobilosti (seznam publikací)

1 - Beneš, V.: Inteligence means to be rewarded. In: Proc. of 5th International PhD Workshop on Systems and Control, Hungary, September 2004, ISBN 963 311359-8, pp. 17 – 22.

Ekštein Kamil Ing. PhD. 770530/2011 Česká republika
člen řešitelského týmu
odborný asistent
377 632 491 kekstein@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj metod akustické analýzy, vývoj a implementace metod segmentace a značkování segmentů, vývoj klasifikátoru, vedení studentů při výzkumu

Prokázání odborné způsobilosti (seznam publikací)

1 - Ekštein, K.; Mouček, R.: Time-Domain Structural Analysis of Speech. In: Proceedings of CICLing 2003, Mexico City, Feb 2003
2 - Ekštein, K.; Matoušek, V.; Mouček, R.: Hybrid analytic/ANN-based acoustic-phonetic decoding. In: Elektronische Sprachsignalverarbeitung. Tagungsband der 14. Konferenz, Karlsruhe, w.e.b. Universitätsverlag & Buchhandel, Dresden, 2003, S. 14-21
3 - Ekštein, K.; Matoušek, V.; Pavelka, T.: Automatic segmentation and labeling of speech signal. In: Elektronische Sprachsignalverarbeitung, Tagungsband der 13. Konferenz, Dresden, w.e.b. Universitätsverlag & Buchhandel, 2002, S. 75-81
4 - Ekštein, K.; Pavelka, T.: LINGVO/LASER: Prototyping concept of dialogue information system with spreading knowledge. In: Proceedings of the 1st International Workshop on Natural Language Understanding and Cognitive Sciences, NLUCS 2004, INSTICC Press, Porto, Portugalsko, 2004
5 - Ekštein, K.; Hitzenberger, L.; Klečková, J.; Krutišová, J.; Kubišta, J.; Matoušek, V.; Mouček, R.; Tauše r, K.: Novel communication concepts for municipal information services. In: SoftCOM 2003 : International conference on software, telecommunications and computer networks. University of Split, 2003. S. 705-709

Fiala Dalibor Ing. 800323/5845 Česká republika
člen řešitelského týmu
výzkumný pracovník
377632479 dalfa@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

Výzkum a implementace prostředků pro analýzu webovských stránek

Prokázání odborné způsobilosti (seznam publikací)

1 - Tesar R., Fiala D., Rousselot F., Jezek K.: A comparison of two algorithms for discovering repeated word sequences, WIT Transactions on Information and Communication Technologies, Vol. 35, pp. 121 - 131, 2005
2 - Belaid A., Alusse A., Rangoni Y., Cecotti H., Farah F., Gagean N., Fiala D., Rousselot F., Vigne H.: Document retro-conversion for personalized electronic reedition, IWDA'05, Calcutta, 2005.

3 - Fiala D., Ježek K.: Retrieving Citations on the Web, Proc. ICKED'04, pp. 481 - 488, Porto, Portugalsko, 2004.

Klečková Jana doc. Dr. Ing. 496108/095 Česká republika
člen řešitelského týmu
docentka
377 632 421 kleckova@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj a implementace metod prozodické analýzy řečového signálu, hodnocení dílčích výsledků, vedení a výchova doktorandů

Prokázání odborné způsobilosti (seznam publikací)

1 - Klečková J., Krutišová J., Matoušek V.: Speaker`s Styles Determinatory by Using Prosody Characteristics. In: Proc. of the 9th International Workshop on Systems, Signals and Image Processing, World Scientific Publ., Ltd., Manchester, November 2002, pp. 370 – 375
2 - Klečková J., Krutišová J., Matoušek V.: Important Prosody Characteristics for Spontaneous Speech Recognition. In: Proceedings of the Int. Conf. on Neural Information Processing ICONIP `02, Nanyang Technological University Publ., Singapore, November 2002, pp. 455 – 461
3 - Klečková, J.: Speech production: Phonetic encoding of real and non-words. In: Text, Speech and Dialogue . Berlin: Springer, 2003, pp. 281-286
4 - Klečková, J.: An investigation of the speech production. In: Models and Analysis of Vocal Emissions for Biomedical Applications. Firenze: Firenze University Press, 2003, pp. 91-93
5 - Klečková, J.: Using nonverbal communication for speaker recognition. In: Proceedings of the NCEI 2004, AVT Technology Park, Auckland Univ. Of Technology, December 2004, ISBN 0-476-01282-1, pp. 37 – 38

Konopík Miloslav Ing. 810326/1782 Česká republika
člen řešitelského týmu
výzkumný pracovník
377 632 491 konopik@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj a implementace metod lingvistické analýzy promluv i textu, porozumění a reprezentace obsahu promluvy či textu, konverze textových a řečových souborů

Prokázání odborné způsobilosti (seznam publikací)

1 - Konopík, M.: Metody rozpoznávání přirozeného jazyka – návrh a implementace modulu sémantické interpretace. Diplomová práce, Západočeská univerzita v Plzni, 2004
2 - Konopík, M., Mouček, R.: An alternative way of semantic interpretation. In: Matoušek, V., Mautner, P., Pavelka T.(Eds.): Text, Speech and Dialogue; Proceedings of the 8th Int. Conference, TSD 2005, Karlovy Vary, September 2005, pp.348 - 355

Krutišová Jana Ing. 595516/0046 Česká republika
člen řešitelského týmu
odborná asistentka, tajemnice katedry
377 632 413 krutisova@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj a implementace metod analýzy spontánních promluv, vedení studentů při výzkumu

Prokázání odborné způsobilosti (seznam publikací)

1 - Klečková J., Krutišová J., Matoušek V.: Speaker`s Styles Determinatory by Using Prosody Characteristics. In: Proc. of the 9th International Workshop on Systems, Signals and Image Processing, World Scientific Publ., Ltd., Manchester, November 2002, pp. 370 – 375
2 - Klečková J., Krutišová J., Matoušek V.: Important Prosody Characteristics for Spontaneous Speech Recognition. In: Proceedings of the Int. Conf. on Neural Information Processing ICONIP `02, Nanyang Technological University Publ., Singapore, November 2002, pp. 455 – 461
3 - Ekštein, K.; Hitzenberger, L.; Klečková, J.; Krutišová, J.; Kubišta, J.; Matoušek, V.; Mouček, R.; Taušer, K.: Novel communication concepts for municipal information services. In: SoftCOM 2003 : International conference on software, telecommunications and computer networks. University of Split, 2003. S. 705-709
4 - Klečková, J.; Krutišová, J.: Some experiments in the Czech spontaneous speech recognition domain. In: Models and Analysis of Vocal Emissions for Biomedical Applications. Firenze: Firenze University Press, 2003, pp. 211-213
5 - Klečková, J.; Krutišová, J.: Can data mining aid automatic spontaneous speech processing? In: Proceedings of the NCEI 2004, AVT Technology Park, Auckland Univ. of Technology, December 2004, pp. 43 – 44

Matoušek Václav prof. Ing. CSc. 480613/108 Česká republika
člen řešitelského týmu
profesor, vedoucí oddělení
377 632 471 matousek@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vedení a koordinace dílčích výzkumných prací, vedení a výchova doktorandů

Prokázání odborné způsobilosti (seznam publikací)

1 - Matoušek V.: Dialogsysteme in eCommerce für Behinderte. In: Fellbaum K. (Hrsg.): Elektronische Sprachsignalverarbeitung (ESSV 2000), Band 20, Tagungsband der elften Konferenz in Cottbus, v.e.b. Universitätsverlag Dresden, September 2000, pp. 231 – 239
2 - Schwarz J., Matoušek V.: Automatic Analysis of Real Dialogues and Generation of Training Corpora. In: Proceedings of the Int. Conference EUROSPEECH 2001, Aalborg, Danmark, September 2001, Volume 4, pp. 2201 – 2204
3 - Freiheit A., Lehner F., Matoušek V.: VoiceXML – Programmierung und Applikationen. VDE Verlag GmbH, Berlin, Offenbach, 2003
4 - Matoušek, V.; Ekštein, K., Pavelka T.: System of an automatic speech recognition and speech understanding. In: Fellbaum K. (Ed.). Elektronische Sprachsignalverarbeitung, Proc. of 15th International Conference on Speech Signal Processing (ESSV 2004), Cottbus, Germany, September 2004, pp. 61 – 68
5 - Matoušek, V.; Kopeček, I.: Formal model of a dialogue. In: Vích, R. (Ed.): Electronic Speech Signal Processing; Proceedings of the 16th Conference joined with the 15th Czech-German Workshop „Speech Processing“, Prague; TUD Press, September 2005, pp. 132 – 142

Mautner Pavel Ing. PhD. 650522/2592 Česká republika
člen řešitelského týmu
odborný asistent
377 632 441 mautner@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj metod analýzy a rozpoznávání spontánních promluv, vedení studentů při výzkumných pracích

Prokázání odborné způsobilosti (seznam publikací)

1 - Mautner P., Rohlík O., Matoušek V., Kempf J.: Fast Signature Verification without a Special Tablet. In: Proceedings of the Int. Conf. IWSSIP ´02, Manchester, November 2002, pp. 496 – 500
2 - Mautner, P., Rohlík, O., Matoušek, V., Kempf, J.: Signature Verification Using Unsupervised Learned Neural Network. Proceedings of the 1st IAPR-TC3 Workshop, Florence, Italy, 2003, pp. 71-75
3 - Mautner, P., Rohlík, O., Matoušek, V., Kempf, J.: Signature Verification Using Self-organizing Feature Map. Proceedings of the 2nd International Conference on Computational Inteligence, Robotics and Autonomous Systems (CIRAS), Singapore, 2003, ps080305.pdf
4 - Rohlík, O.; Mautner, P.; Matoušek, V.; Kempf, J.: HMM based handwritten text recognition using biometrical data acquisition pen. In: CIRA2003 [elektronický zdroj] : Proceedings 2003 IEEE international symposium on computational intelligence in robotics and automation. Kobe: University of Kobe, 2003, S. 950-953
5 - Mautner, P.; Maršálek, T.; Rohlík, O.; Matoušek, V.: Comparison of ART-2 and SOFM Based Neural Network Verifiers. In: Proceedings of the 1st Int. Workshop on Artificial Neural Networks (ANNs 2004) , INSTICC Press, Setúbal, Portugal, August 2004, pp. 43 – 50

Mouček Roman Ing. PhD. 760707/2000 Česká republika
člen řešitelského týmu
odborný asistent
377 632 441 moucek@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj a implementace metod analýzy a zpracování větné sémantiky, analýza větných struktur, vedení studentů

Prokázání odborné způsobilosti (seznam publikací)

1 - Matoušek V., Mouček R., Taušer K.: Ein Dialogsystem für blinde und sehbehinderte Programmierer. In: Hess W., Stöber K. (Hrsg.): Elektronische Sprachsignalverarbeitung (ESSV 2001), Band 22, Tagungsband der zwölften Konferenz in Bonn, v.e.b. Universitätsverlag Dresden, September 2001, pp. 175 – 182
2 - Mouček, R., Ekštein, K.: Corpus Construction within Linguistic Module of City Information Dialogue System. In: Proceedings of CICLing 2003, Mexico City, Feb 2003
3 - Mouček, R.: Semantic Hierachies in Dialogue Systems. In: Proc. of the 2004 WSEAS Conferences, Izmir, September 2004 (ICOSSIP 2004), CD ROM
4 - Mouček, R.: Semantics in Computer Dialogue Systems. In: Proc. of the 6th Int. Scientific Conference on Electronic Computers and Informatics, Košice-Herlany, September 2004, VIENALA Press, Košice, 2004, pp. 38 – 43
5 - Mouček, R.; Andrš, D.: Theory of microsituations in spoken language. In: Proc. of 5th International PhD Workshop on Systems and Control, Hungary, September 2004, pp. 83 - 86

Pavelka Tomáš Ing. 790918/2083 Česká republika
člen řešitelského týmu
výzkumný pracovník
377 632 491 tpavelka@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

vývoj a implementace metod analýzy a konkatenace segmentů, syntaktická kontrola rozpoznávaných vět

Prokázání odborné způsobilosti (seznam publikací)

1 - Ekštein, K.; Matoušek, V.; Pavelka, T.: Automatic segmentation and labeling of speech signal. In: Elektronische Sprachsignalverarbeitung, Tagungsband der 14. Konferenz, Dresden, w.e.b. Universitätsverlag & Buchhandel, 2003, S. 75 – 81
2 - Brey, T., Pavelka, T.: A speech platform for a bilingual city information system. In: Proceedings of the Int. Conference TSD 2004, LNAI 3206, Springer Verlag, Berlin, Heidelberg, 2004, pp. 571 – 578
3 - Ekštein, K.; Pavelka, T.: LINGVO/LASER: Prototyping concept of dialogue information system with spreading knowledge. In: Proceedings of the 1st International Workshop on Natural Language Understanding and Cognitive Sciences, NLUCS 2004, INSTICC Press, Porto, Portugalsko, 2004
4 - Matoušek, V.; Ekštein, K., Pavelka T.: System of an automatic speech recognition and speech understanding. In: Fellbaum K. (Ed.). Elektronische Sprachsignalverarbeitung, Proc. of 15th International Conference on Speech Signal Processing (ESSV 2004), Cottbus, Germany, September 2004, pp. 61 – 68
5 - Pavelka, T.; Ekštein, K.: Low level performance on continuous speech: humans vs. computers.. In: Proc. of 5th International PhD Workshop on Systems and Control, Hungary, September 2004, pp. 109 – 112

Steinberger Josef Ing. 790918/2127 Česká republika
člen řešitelského týmu
výzkumný pracovník
377 632 479 jstein@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

Výzkum metod automatické sumarizace textu, rezoluce anafor a jejich implementace

Prokázání odborné způsobilosti (seznam publikací)

1 - Steinberger, J., Jezek, K.: Using Latent Semantic Analysis in Text Summarization, ISIM04, pp. 93-100, MARQ Ostrava

2 - Steinberger, J., Jezek, K.: Text Summarization and Singular Value Decomposition, ADVIS 2004, Lecture Notes in Comp.Sc.2457 pp.245-254, Springer-Verlag 2004
3 - Steinberger, J., Ježek, K.: Hodnocení kvality sumarizátorů textů, ZNALOSTI2005, pp.96-107, Stará Lesná, Slovensko 2005
4 - Steinberger, J., Kabadjov, M.A., Poesio M., Sanchez-Graillet, O.: Improving LSA-based Summarization with Anaphora Resolution, EMNLP'05, pp.1-8, Association for Computation Linguistics
5 - Kabadjov, M.A., Poesio, M., Steinberger, J.: Task-Based Evaluation of Anaphora Resolution: The Case of Summarization, RANLP'05 Workshop Crossing Barriers in Text Summarization Research", Borovets, Bulharsko

Tesař Roman Ing. 790930/2379 Česká republika
člen řešitelského týmu
výzkumný pracovník
377632479 romant@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

Výzkum a implementace metod klasifikace, filtrace a vyhledávání www stránek

Prokázání odborné způsobilosti (seznam publikací)

1 - Tesař, R., Ježek, K.: Filtrace webových stránek Suffix Tree frázemi, Proc. of ITAT 2004, Lences R. (ed.), Slovensko, září 2004

2 - Tesař, R., Ježek, K.: Klasifikace Sufix Tree frázemi - srovnání s metodou Itemsets, Proc. ZNALOSTI 2005, pp. 144 - 153, Stará Lesná, Slovensko, 2005

3 - Tesar, R., Fiala, D., Rousselot, F., Jezek, K.: A comparison of two algorithms for discovering repeated word sequences, Data Mining IV 2005, pp. 121-131, Skiathos, Greece, WIT Press, WIT Transaction on Information and Communication Technologies, 2005

Toman Michal Ing. 800704/2054 Česká republika
člen řešitelského týmu
výzkumný pracovník
377632479 mtoman@kiv.zcu.cz

Stěžejní vykonávané činnosti při řešení projektu

Vývoj a implementace prostředků extrakce znalostí z textů v multijazykovém prostředí

Prokázání odborné způsobilosti (seznam publikací)

1 - Jezek, K., Toman, M.: Documents Categorization in Multilingual Environment, ElPub2005, pp.97-104, Leuven, Belgium 2005, Peeters Publishing
2 - Toman M., Jezek K.: Modifikace bayesovského disambiguátoru, Znalosti 2005, pp.306-313, VŠB-Technická univerzita Ostrava 2005
3 - Toman M., Jezek K.: Klasifikace multilinguálních korpusů s využitím tezauru EuroWordNet, Proceedings of ITAT, Slovensko 2004

4.2.2.3. - Uchazeč (základní údaje uchazeče - právního subjektu příslušného pracoviště)
"Západočeská univerzita v Plzni"

Role uchazeče	000 - příjemce - koordinátor
IČ	49777513
Obchodní jméno - Název	**Západočeská univerzita v Plzni**
Právní forma subjektu	VVS
Adresa sídla – Ulice	Univerzitní 8
Adresa sídla – Místo	Plzeň
Adresa sídla – PSČ	30614
Adresa sídla – Stát	CZ
Telefonické spojení	377 631 111
Bankovní spojení organizace	
Kód banky	0100
Název banky	Komerční banka, a.s., Plzeň
Číslo účtu	4811530257
Specifický symbol	
DIČ	CZ49777513
Zkratka názvu organizace	ZČU
WWW adresa	www.zcu.cz
Zápis v obchodním rejstříku	
- vedený kde	
- oddíl	
- vložka	

4.2.2.4. Statutární orgán uchazeče

Průša Josef doc. Ing. CSc. - rektor 377 631 000 rektor@rek.zcu.cz

Razítko:			
Datum:			
Podpis(y):	Průša Josef doc. Ing. CSc. rektor		

Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

4.2.2.1. - Údaje o pracovišti - Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

Název pracoviště

Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

Popis pracoviště podílejícího se na činnosti centra , jeho lokalizace a popis prostor vymezených pro jeho činnost

Centrum zpracování přirozeného jazyka (CZPJ) je umístěno zejména v komplexu dvou velkých místností (laboratoří) Fakulty informatiky MU ve druhém patře budovy B. Součástí centra je též 5 kanceláří docentů a odborných asistentů Centra ve třetím patře budovy B. Centrum je vybaveno dvěma výkonnými servery a 23 pracovními stanicemi, které slouží pro práci členů laboratoře, doktorandů a také pregraduálních studentů. Server spolu s pracovními stanicemi umožňuje zpracování velkých datových souborů (řádově ve stovkách GB).
Základní vybavení představuje:
* 1 server - 2 CPU Intel Xeon 2.2GHz, 4 GB RAM, 500 GB disk, OS Linux
* 1 server - 4 CPU AMD Opteron 2.4GHz, 24 GB RAM, 3 TB disk, OS Linux
* 23 x pracovní stanice na bázi PC, OS Linux
* 1 x scanner
* 1x datový projektor
* 2 laserové tiskárny
Všechny prostory CZPJ jsou vybaveny síťovou infrastrukturou 100/1000 Mbit Ethernet, která je součástí celofakultní sítě. Laboratorní servery jsou připojeny k fakultní síti na úrovni 1 GB spojení. Centrum je rovněž pokryto signálem celofakultní bezdrátové sítě. Členové centra též mohu využívat další zařízení Fakulty informatiky jako například knihovnu, velkokapacitní kopírovací stroje a další.

Dosavadní výzkumná činnost

Činnost centra se skládá ze základního a aplikovaného výzkumu v oblasti zpracování přirozeného jazyka. Výsledky v aplikovaném výzkumu vyústily v celou řadu systémů, pět nejvýznamnějších vztahujících se k podávanému projektu je uvedeno v následujícím seznamu:
* MANATEE: Korpusový manažer Manatee umožňuje zpracovávat rozsáhlé korpusy s velikostí přes dvě miliardy slovních tvarů. Je jazykově nezávislý a dokáže rychle vyhledávat a počítat různé statistiky. Systém obsahuje grafické uživatelské rozhraní Bonito a aplikační programové rozhraní (API). Systém se rutinně používá na řadě pracovišť v ČR i po celém světě (Slovensko, Německo, Velká Británie, Rusko, Itálie, USA).
* AJKA: Morfologický analyzátor a generátor AJKA pro zadané české slovo určí vzor, základní tvar, všechny gramatické kategorie, segmentaci a případné slovotvorné vazby na jiná slova. Existuje ve formě spustitelného programu, ale stejná funkcionalita je přístupná i přes API pro programovací jazyky C, Perl, Python, Ruby. K dispozici je též webové rozhraní(http://nlp.fi.muni.cz/projekty/wwwajka/). Přestože byl analyzátor Ajka primárně vyvíjen pro češtinu, existuje v současné době i morfologická databáze pro slovenštinu a experimentálně bylo ověřeno použití pro jiné jazyky (francouzština, holandština, španělština). Systém je rutinně používán pro značkování korpusů.
* SYNT: Tabulkový syntaktický analyzátor SYNT pro češtinu je budován na základě formálních metapravidel. Je testován, aby mohl sloužit v rámci normálního translačního algoritmu (NTA), který vstupním českým větám dovede přiřadit jejich sémantické reprezentace v podobě formulí lambda kalkulu. K dispozici je implementace knihovny pro vlastní syntaktický analyzátor SYNT a navazující modul NTA. Systém SYNT je budován jako jazykově nezávislý, je optimalizován pro rozsáhlé a velmi nejednoznačné gramatiky (řádově čítající tisíce až desetitisíce odvozovacích pravidel).
* VisDic: Nástroj pro prohlížení a editaci elektronických lexikálních databází uložených ve formátu XML s důrazem na vícejazyčná data s velkým počtem vazeb. VisDic a odpovídající XML formát lexikální databáze se používá jako standardní systém pro budování a zpracování národních WordNetů v rámci EU projektu Balkanet (IST 2000 29388) a je využíván i pro další jazyky (např. angličtina, ruština, němčina).
* Algebraické modely dialogových systémů a uživatele: Na bázi Mealyho automatů byl vytvořen model dialogových systémů, model osobnosti uživatele a jeho interpretace pro modelování emočních stavů uživatele. Teorie byla publikována na mezinárodních konferencích (AIA 2004, ICCC 2003, User Modeling 2003, ICON 2002, User Modeling 2001). Byl vytvořen model komunikace dialogový systém – uživatel, kombinující stochastický (uživatel) a deterministický model a teorie byla uplatněna pro simulování chování osobnosti uživatele.

Popis výzkumné činnosti pracoviště a jeho vybaveni

Výzkumná činnost CZPJ bude zasahovat do všech čtyřech dílčích cílů navrhovaného projektu. Konkrétně bude se zaměříme na:

* korpus syntaktických stromů,
* korpus vzorových přepisů vybraných vět na jejich sémantické reprezentace,
* doplnění morfologického značkovače o robustní hádací proceduru,
* algoritmy a programy pro přiřazování logických formulí větám z volného textu,
* algoritmy pro vyvozování nových znalostí z informací získaných z volného textu,
* vytvoření nástrojů a metod syntaktické analýzy textů v přirozeném jazyce,
* uživatelského rozhraní pro práci se sémantickým webem

* zpracování a ověření algoritmu logické analýzy české věty za účelem reprezentace znalostí,
* zpřístupnění webu širší skupině uživatelů, včetně uživatelů handicapovaných,
* využití struktury i sémantiky webových stránek k extrahování znalostí,
* komunikace s www prostředky přirozeného jazyka.

Současné materiálně technické podmínky CZPJ, jak je popsáno v předchozí kapitole 4.2.2.1.2., spolu s novými investicemi plánovanými v tomto podávaném projektu budou vhodné pro dosažení stanovených dílčích cílů.

Vyjádření k duplicitě řešení

Fakulta informatiky MU je zapojena do projektu „Inteligentní modely, algoritmy, metody a nástroje pro vytváření sémantického webu“ (Projekt programu Informační společnost tématického programu II Národního programu výzkumu v České republice Identifikační kód: 1ET100300419). Cílem projektu je vyzkoušet a vyvinout teoretické základy sémantického webu a jde tedy o základní výzkum. Vybraní pracovníci FI MU pracují v rámci Centra základního výzkumu „Integrované centrum počítačového zpracování přirozeného jazyka“ (MŠMT LC536).
V současné době je také podáván návrh výzkumného záměru s názvem „Pokročilé zpracování digitálních dat“, ve kterém půjde o základní výzkum zejména v oblasti vyhledávání různě strukturovaných dat. Někteří členové CZPJ se na uvedených projektech přímo podílejí. V tomto navrhovaném projektu hodláme navázat na teoretické výsledky z výše uvedených projektů.
Navrhovaný projekt bude též navazovat na probíhající, resp. ukončený grant Grantové agentrury ČR: „Překlad českých vět na konstrukce transparentní intenzionální logiky“ (č.201/05/2781)
a „Velké jazykové korpusy a jejich automatická analýza“ (č. 405/03/0913)

4.2.2.2. Řešitelský tým - Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

Bártek Luděk Mgr. 720108/3791 Česká republika
člen řešitelského týmu
Horák Aleš Mgr. Ph.D. 740901/4250 Česká republika
člen řešitelského týmu
Kopeček Ivan doc. RNDr. CSc. 490303/075 Česká republika
člen řešitelského týmu
Pomikálek Jan Mgr. 791009/0419 Česká republika
člen řešitelského týmu
Rychlý Pavel Mgr. Ph.D. 730123/5359 Česká republika
člen řešitelského týmu
Sojka Petr RNDr. Ph.D. 630917/1000 Česká republika
člen řešitelského týmu
Pala Karel doc. PhDr. CSc. 390615/416 Česká republika
spoluřešitel

Bártek Luděk Mgr. 720108/3791 Česká republika
člen řešitelského týmu
asistent
549 49 3215 bar@fi.muni.cz

Stěžejní vykonávané činnosti při řešení projektu

Návrh, vývoj a implementace metod pro generování dialogových rozhraní.

Prokázání odborné způsobilosti (seznam publikací)

1 - Kopeček, I.; Bártek, L.: Adapting Web-Based Educational Systems for the Visually Impaired. In: CIAH 2005, Proceedings of International Workshop. Salzburg : 2005. pp. 11-16.
2 - Bártek, L.: Improvements in a Dialogue Interface for Library System. In: Applied Informatics (AI 2003). Calgary (Alberta, CA) : IASTED, 2003. pp. 162-165, ISBN 0-88986-341-5.
3 - Bártek, L.: Automatic Generation of Dialogue Interaces for Web-Based Applications. In: Text, Speech and Dialogue. Berlin Heidelberg : Springer-Verlag, 2001. pp. 443-449. LNAI 2166. ISBN 3-540-42557-8.
4 - Bártek, L.: Phonetic Corpus Based on PHC Format. In: Proceedings ISM'99. Ostrava (CZ) : MARQ, Tomas Hruska, 1999. pp. 169-175. ISBN 80-85988-31-3.
5 - Luděk, B.: User Interface of the Spoken Language Corpus CLAP, presented at COCOSDA'99 meeting, Budapest, Hungary, 1999, (http://www.slt.atr.co.jp/cocosda/99.html)

Horák Aleš Mgr. Ph.D. 740901/4250 Česká republika
člen řešitelského týmu
odborný asistent
549 49 4377 hales@fi.muni.cz

Stěžejní vykonávané činnosti při řešení projektu

Výzkum a organizace výzkumné činnosti v oblasti syntaktické analýzy, logické analýzy věty, reprezentace znalostí a vyvozování znalostí.

Prokázání odborné způsobilosti (seznam publikací)

1 - Horák, Aleš - Kadlec, Vladimír. New Meta-grammar Constructs in Czech Language Parser synt. Lecture Notes in Artificial Intelligence, Berlin : Springer-Verlag, 3658/2005, 1, od s. 85-92, 8 s. ISSN 0302-9743. 2005.
2 - Hlaváčková, Dana - Horák, Aleš. VerbaLex - New Comprehensive Lexicon of Verb Valencies for Czech. In Computer Treatment of Slavic and East European Languages. Bratislava, Slovakia : Slovenský národný korpus, 2006.
3 - Horák, Aleš - Pala, Karel. Lexicons in TIL and Verb Valency Frames. In Proceedings of the International Conference on Communications in Computing (CIC 2004). Las Vegas, Nevada, USA : CSREA Press, 2004.
4 - Horák, Aleš - Smrž, Pavel. New Features of Wordnet Editor VisDic. Romanian Journal of Information Science and Technology, Romanian Academy, 7, 1-2, od s. 201-214, 14 s. ISSN 1453-8245. 2004.
5 - Horák, Aleš - Smrž, Pavel. Best Analysis Selection in Inflectional Languages. In Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan : The Association for Computational Linguistics and Chinese Language Processing, 2002. s. 363-368. ISBN 1-55860-894-X.

Kopeček Ivan doc. RNDr. CSc. 490303/075 Česká republika
člen řešitelského týmu
docent
549 49 3861 kopecek@fi.muni.cz

Stěžejní vykonávané činnosti při řešení projektu

výzkum v oblasti modelování dialogových systémů a interakce člověk-počítač, vedení kolektivu Laboratoře řeči a dialogu, pravidelná kontrola a hodnocení výsledků, vedení a výchova doktorandů

Prokázání odborné způsobilosti (seznam publikací)

1 - Kopeček, Ivan. Optimal Structures of Speech Menus in Speech Based Hypertexts and Dialogue Systems. In Proceeedings of the 10th International Conference Speech and Computer, Patras, Greece, pp. 719-723, 2005, ISBN 5-7452-0110-x.

2 - Kopeček, Ivan. Dialogues for Agent Communication. In Proceedings of the International Conference on Communications in Computing CIC'04, Las Vegas : CSREA Press, 2004. pp. 262-266, ISBN 1-932415-36-X.
3 - Kopeček, Ivan - Batůšek, Robert - Kučera, Antonín. On Homogeneous Segments. In Text, Speech and Dialogue - 6th International Conference, Proceedings. Berlin : Springer Verlag, 2003. LNAI 2807. pp. 152-157. ISBN 0302-9743.
4 - Kopeček, Ivan - Novotný, Miroslav. On Equations Including Strings. Fundamenta Informaticae, IOS Press, pp. 391-398, ISSN 0169-2968. 2002.
5 - Kopeček, Ivan. Algebraic models of Speech Segment Databases. In Proceedings of the international conference on Text, Speech and Dialogue, Springer Verlag, LNAI 2448. pp. 208-213, 2001, ISBN 3-540-42557-8.

Pomikálek Jan Mgr. 791009/0419 Česká republika
člen řešitelského týmu
odborný pracovník
549 49 1864 xpomikal@fi.muni.cz

Stěžejní vykonávané činnosti při řešení projektu

Vývoj technik a nástrojů v oblasti zpracování korpusových dat, sumarizace textů, extrakce termínů a podobnosti dokumentů.

Prokázání odborné způsobilosti (seznam publikací)

1 - Adam Kilgarriff, Chu-Ren Huang, Michael Rundell, Jan Pomikalek, Pavel Rychly, Simon Smith, David Tugwell, Elaine Uí Dhonnchadha. Word sketches for Irish and
Chinese. In Proceedings from the Corpus Linguistics Conference Series, ISSN 1747-9398, 2005.

Rychlý Pavel Mgr. Ph.D. 730123/5359 Česká republika
člen řešitelského týmu
odborný asistent
549 49 6399 pary@fi.muni.cz

Stěžejní vykonávané činnosti při řešení projektu

Organizace a vývoj v oblasti zpracování korpusových dat, ontologie a reprezentace znalostí.

Prokázání odborné způsobilosti (seznam publikací)

1 - Kilgarriff, Adam - Rychlý, Pavel - Smrž, Pavel - Tugwell, David. The Sketch Engine. In Proceedings of the Eleventh EURALEX International Congress. Lorient, France : Universite de Bretagne-Sud, 2004. od s. 105-116, 12 s. ISBN 2952245703.
2 - Rychlý, Pavel - Smrž, Pavel. Manatee, Bonito and Word Sketches for Czech. In Proceedings of the Second International Conference on Corpus Linguisitcs. Saint-Petersburg : Saint-Petersburg State University Press, 2004. od s. 124-132, 9 s. ISBN 5-288-03531-8.
3 - Pala, Karel - Rychlý, Pavel - Smrž, Pavel. Text Corpus with Errors. In Text, Speech and Dialogue: Sixth International Conference, TSD 2003. Berlin : Springer Verlag, 2003. od s. 90-97, 8 s. LNAI 2807. ISBN 3-540-200-24-X.
4 - Rychlý, Pavel. GCQP -- Multiplatform Graphical User Interface to the CQP corpus manager. In Proceedings of the Ninth EURALEX International Congress. Stuttgart : Institut fur Maschinelle Sprachverarbeitung, 2000. s. 149-154. ISBN 3-00-006574-1.
5 - Smrž, Pavel - Rychlý, Pavel. Finding Semantically Related Words in Large Corpora. In Text, Speech and Dialogue, 4th International Conference, TSD 2001. Berlin : Springer-Verlag, 2001. s. 108-115. LNAI 2166. ISBN 3-540-42557-8.

Sojka Petr RNDr. Ph.D. 630917/1000 Česká republika
člen řešitelského týmu
odborný asistent
549496966 sojka@fi.muni.cz

Stěžejní vykonávané činnosti při řešení projektu

Výzkum a vývoj v oblasti morfologická desambiguace, segmentace řeči a textu technikou přebíjejících vzorů, vizualizační algoritmy a postupy (VisualBrowser), extrakce textů z dokumentů a textových informačních systémů.

Prokázání odborné způsobilosti (seznam publikací)

1 - Sojka, Petr - Choi, Key-Sun - Fellbaum, Christiane - Vossen, Piek (Eds).: Proceedings of the Third International WordNet Conference, GWC 2006, Seogwipo, Korea, January 22-26, 2006. Brno : Masaryk University, 2006. 362 s. ISBN 80-210-3915-9.
2 - Sojka, Petr. From Scanned Image to Knowledge Sharing. In Proceedings of I-KNOW'05. Graz, Austria : Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co., 2005. pp. 664-672, ISBN 0948-6968.

3 - Holeček, Jan - Sojka, Petr. Animations in pdfTeX-generated PDF. Lecture Notes in Computer Science, Berlin, Heidelberg : Springer-Verlag, 2004, 3130, pp. 179--191, ISSN 0302-9743.
4 - Sojka, Petr - Antoš, David. Context Sensitive Pattern Based Segmentation: A Thai Challenge. In Proceedings of EACL 2003 workshop Computational Linguistics for South Asian Languages -- Expanding Synergies with Europe. Budapest : Association for Computational Linguistics, 2003. pp. 65-72, ISBN 1-932432-02-7.
5 - Sojka, Petr. Competing Patterns for Language Engineering. In Proceedings of Third International Workshop on Text, Speech and Dialogue, TSD 2000. Heidelberg : Springer-Verlag, 2000. pp. 157-162. Lecture Notes in Artificial Intelligence 1902. ISBN 3-540-41042-2.

Pala Karel doc. PhDr. CSc. 390615/416 Česká republika
spoluřešitel
vedoucí Centra zpracování přirozeného jazyka
549 49 5616 pala@fi.muni.cz

Stěžejní vykonávané činnosti při řešení projektu

Vedení a koordinace prací na projektu.

Prokázání odborné způsobilosti (seznam publikací)

1 - Horák, Aleš - Pala, Karel - Rambousek, Adam - Oliva, Karel. Preparation of Czech Lexical Database with the PRALED DEBII Platform Tool. In Grammar & Corpora. 2006. vyd. Praha : Ústav pro jazyk český Akademie věd ČR, 2006.
2 - Pala, Karel - Sedláček, Radek. Enriching WordNet with Derivational Subnets. In Sedláček, Radek. Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing CICLING 2005. Berlin Heidelberg New York : Springer Verlag, 2005. od s. 305-311, 7 s. ISBN 3-540-24523-5.
3 - Pala, Karel - Smrž, Pavel. Building Czech Wordnet. Romanian Journal of Information Science and Technology, Romanian Academy, 7, 1-2, od s. 79-88, 10 s. ISSN 1453-8245. 2004.
4 - Pala, Karel - Smrž, Pavel. Ontologies, Types and Verb Frames. In Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2004. Calgary, Alberta, Canada : ACTA Press, 2004. od s. 874-877, 4 s. ISBN 0-88986-404-7.
5 - Sojka, Petr - Pala, Karel - Smrž, Pavel - Fellbaum, Christine - Vossen, Piek. Proceedings of the Second International WordNet Conference, GWC 2004, Brno, Czech Republic, January 20-23, 2004. Edited by Sojka P., Pala K., Smrž P., Fellbaum Ch., Vossen P. prvni. Brno : Masaryk University, 2004. 370 s. GWC Proceedings. Second International WordNet Conference, GWC 2004, Brno, Czech Republic, January 20--23, 2004, Proceedings. ISBN 80-210-3302-9.

4.2.2.3. - Uchazeč (základní údaje uchazeče - právního subjektu příslušného pracoviště)
"Masarykova univerzita"

Role uchazeče	020 - spolupříjemce
IČ	00216224
Obchodní jméno - Název	**Masarykova univerzita**
Právní forma subjektu	VVS
Adresa sídla – Ulice	Žerotínovo nám. 9
Adresa sídla – Místo	Brno
Adresa sídla – PSČ	60177
Adresa sídla – Stát	CZ
Telefonické spojení	549 491 1111
Bankovní spojení organizace	
Kód banky	0100
Název banky	Komerční banka Brno-město
Číslo účtu	85636621
Specifický symbol	
DIČ	CZ00216224
Zkratka názvu organizace	MU
WWW adresa	www.muni.cz
Zápis v obchodním rejstříku	
- vedený kde	
- oddíl	
- vložka	

4.2.2.4. Statutární orgán uchazeče
Fiala Petr prof. PhDr. Ph.D. - rektor Masarykovy univerzity 549 49 1001 rektor@muni.cz

Razítko:			
Datum:			
Podpis(y):	Fiala Petr prof. PhDr. Ph.D. rektor Masarykovy univerzity		

4.3. FINANČNÍ PLÁN

4.3.1. Uznané náklady souhrnně za projekt

NÁKLADY		2006		2007		2008		2009		2010		2011		CELKEM	
		UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč
F1	Osobní Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	2403	2231	4653	4364	4787	4466	4815	4487	4842	4507	0	0	21500	20055
F1.1	Mzdy a platy Mzdy a platy	1703	1603	3374	3174	3480	3260	3500	3275	3520	3290	0	0	15577	14602
F1.2	Dohody Úhrada dohod o pracích konaných mimo pracovní poměr	50	15	30	15	20	0	20	0	20	0	0	0	140	30
F1.3	Povinné zákonné odvody Povinné zákonné odvody včetně případného přídělu do FKSP	650	613	1249	1175	1287	1206	1295	1212	1302	1217	0	0	5783	5423
F2	Pořízení majetku Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	1169	715	530	210	0	0	0	0	0	0	0	0	1699	925
F3	Provoz a údržba Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	80	45	160	40	160	40	160	40	160	40	0	0	720	205
F4	Další provozní Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	80	45	160	40	160	40	160	40	140	40	0	0	700	205
F5	Služby Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	10	0	20	0	30	0	30	0	20	0	0	0	110	0
F6	Výsledky Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	10	0	50	0	100	0	100	0	100	0	0	0	360	0
F7	Cestovné Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	150	77	320	139	580	171	580	178	610	185	0	0	2240	750
F8	Doplňkové Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	250	0	500	0	500	0	500	0	500	0	0	0	2250	0
F9	NÁKLADY CELKEM NÁKLADY CELKEM	4152	3113	6393	4793	6317	4717	6345	4745	6372	4772	0	0	29579	22140
F9A	z toho běžné náklady z toho běžné náklady	2983	2398	5863	4583	6317	4717	6345	4745	6372	4772	0	0	27880	21215

ZDROJE		2006	2007	2008	2009	2010	2011	CELKEM
ZD	Dotace Z dotace na projekt	3113	4793	4717	4745	4772	0	22140
ZO	Ostatní veřejné zdroje Ostatní veřejné zdroje	0	0	0	0	0	0	0
ZN	Neveřejné zdroje Neveřejné zdroje	1039	1600	1600	1600	1600	0	7439
ZC	ZDROJE CELKEM ZDROJE CELKEM	4152	6393	6317	6345	6372	0	29579

4.3.1. Nákladové tabulky uchazeče

4.3.1. Tabulka uznaných nákladů - Západočeská univerzita v Plzni

NÁKLADY		2006		2007		2008		2009		2010		2011		CELKEM	
		UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč
F1	Osobní Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	1491	1476	2907	2892	2965	2945	2965	2945	2965	2945	0	0	13293	13203
F1.1	Mzdy a platy Mzdy a platy	1066	1066	2100	2100	2150	2150	2150	2150	2150	2150	0	0	9616	9616
F1.2	Dohody Úhrada dohod o pracích konaných mimo pracovní poměr	30	15	30	15	20	0	20	0	20	0	0	0	120	30
F1.3	Povinné zákonné odvody Povinné zákonné odvody včetně případného přídělu do FKSP	395	395	777	777	795	795	795	795	795	795	0	0	3557	3557
F2	Pořízení majetku Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	1094	640	320	0	0	0	0	0	0	0	0	0	1414	640
F3	Provoz a údržba Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	50	25	100	20	100	0	100	0	100	0	0	0	450	45
F4	Další provozní Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	50	25	100	20	100	0	100	0	80	0	0	0	430	45
F5	Služby Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	10	0	20	0	30	0	30	0	20	0	0	0	110	0
F6	Výsledky Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	10	0	50	0	100	0	100	0	100	0	0	0	360	0
F7	Cestovné Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	100	50	200	65	400	50	400	50	430	50	0	0	1530	265
F8	Doplňkové Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	150	0	300	0	300	0	300	0	300	0	0	0	1350	0
F9	NÁKLADY CELKEM NÁKLADY CELKEM	2955	2216	3997	2997	3995	2995	3995	2995	3995	2995	0	0	18937	14198
F9A	z toho běžné náklady z toho běžné náklady	1861	1576	3677	2997	3995	2995	3995	2995	3995	2995	0	0	17523	13558

ZDROJE		2006	2007	2008	2009	2010	2011	CELKEM
ZD	Dotace Z dotace na projekt	2216	2997	2995	2995	2995	0	14198
ZO	Ostatní veřejné zdroje Ostatní veřejné zdroje	0	0	0	0	0	0	0
ZN	Neveřejné zdroje Neveřejné zdroje	739	1000	1000	1000	1000	0	4739
ZC	ZDROJE CELKEM ZDROJE CELKEM	2955	3997	3995	3995	3995	0	18937

4.3.1. Nákladové tabulky uchazeče

4.3.1. Tabulka uznaných nákladů - Masarykova univerzita

NÁKLADY		2006		2007		2008		2009		2010		2011		CELKEM	
		UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč
F1	Osobní Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	912	755	1746	1472	1822	1521	1850	1542	1877	1562	0	0	8207	6852
F1.1	Mzdy a platy Mzdy a platy	637	537	1274	1074	1330	1110	1350	1125	1370	1140	0	0	5961	4986
F1.2	Dohody Úhrada dohod o pracích konaných mimo pracovní poměr	20	0	0	0	0	0	0	0	0	0	0	0	20	0
F1.3	Povinné zákonné odvody Povinné zákonné odvody včetně případného přídělu do FKSP	255	218	472	398	492	411	500	417	507	422	0	0	2226	1866
F2	Pořízení majetku Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	75	75	210	210	0	0	0	0	0	0	0	0	285	285
F3	Provoz a údržba Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	30	20	60	20	60	40	60	40	60	40	0	0	270	160
F4	Další provozní Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	30	20	60	20	60	40	60	40	60	40	0	0	270	160
F5	Služby Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F6	Výsledky Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F7	Cestovné Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	50	27	120	74	180	121	180	128	180	135	0	0	710	485
F8	Doplňkové Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	100	0	200	0	200	0	200	0	200	0	0	0	900	0
F9	NÁKLADY CELKEM NÁKLADY CELKEM	1197	897	2396	1796	2322	1722	2350	1750	2377	1777	0	0	10642	7942
F9A	z toho běžné náklady z toho běžné náklady	1122	822	2186	1586	2322	1722	2350	1750	2377	1777	0	0	10357	7657

ZDROJE		2006	2007	2008	2009	2010	2011	CELKEM
ZD	Dotace Z dotace na projekt	897	1796	1722	1750	1777	0	7942
ZO	Ostatní veřejné zdroje Ostatní veřejné zdroje	0	0	0	0	0	0	0
ZN	Neveřejné zdroje Neveřejné zdroje	300	600	600	600	600	0	2700
ZC	ZDROJE CELKEM ZDROJE CELKEM	1197	2396	2322	2350	2377	0	10642

4.3.2. Nákladové tabulky pracovišť
Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

4.3.2.1. Náklady na pořízení majetku - Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

Idč	Název	Dodavatel	Rok pořízení	Pořizovací Cena tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE z SRU tis. Kč	
1	Sun Fire V490 Server	INCAD s.r.o., Praha	2006	794	794	490	
2	Diskové pole Sun StorEdge 3310	INCAD s.r.o., Praha	2006	300	300	150	
3	Diskové pole Sun StorEdge 3510 FC Array	INCAD s.r.o., Praha	2007	320	320	0	
	SOUČET ZA ROK		2006		1094	640	
	SOUČET ZA ROK		2007		320	0	

4.3.2.2. Náklady na mzdy a platy - Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

	Jméno	Rodné číslo	1. Rok řešení	% pracovního úvazku	měsíční výše platu nebo mzdy včetně nadtarifních složek tis. Kč	počet měsíců	Roční výše odměny tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE z SRU tis. Kč	
	Ježek Karel doc. Ing. CSc.	420617/110	2006	20	50	6	15	75	75	
	Andrš David Ing.	800204/2653	2006	60	22	6	8	87,2	87,2	
	Beneš Vilém Ing.	800318/2220	2006	30	22	6	4	43,6	43,6	
	Ekštein Kamil Ing. PhD.	770530/2011	2006	30	27	6	10	58,6	58,6	
	Fiala Dalibor Ing.	800323/5845	2006	60	22	6	8	87,2	87,2	
	Klečková Jana doc. Dr. Ing.	496108/095	2006	10	30	6	10	28	28	
	Konopík Miloslav Ing.	810326/1782	2006	60	22	6	8	87,2	87,2	
	Krutišová Jana Ing.	595516/0046	2006	10	25	6	10	25	25	
	Matoušek Václav prof. Ing. CSc.	480613/108	2006	20	50	6	15	75	75	
	Mautner Pavel Ing. PhD.	650522/2592	2006	25	27	6	8	48,5	48,5	
	Mouček Roman Ing. PhD.	760707/2000	2006	25	27	6	8	48,5	48,5	
	Pavelka Tomáš Ing.	790918/2083	2006	100	22	6	8	140	140	
	Steinberger Josef Ing.	790918/2127	2006	60	22	6	8	87,2	87,2	
	Tesař Roman Ing.	790930/2379	2006	60	22	6	8	87,2	87,2	
	Toman Michal Ing.	800704/2054	2006	60	22	6	8	87,2	87,2	
	SOUČET		2006					1066	1066	

4.3.2.3. Náklady na dohody - Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

Idč	Popis činností zajišťovaných v přímé souvislosti s řešením projektu formou dohod o pracích konaných mimo pracovní poměr	1.Rok řešení	Počet hodin	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE z SRU tis. Kč	
1	Pořizování řečových dat	2006	300	30	30	
	SOUČET	2006		30	15	

4.3.2.4. Tabulka uznaných nákladů - Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

NÁKLADY		2006		2007		2008		2009		2010		2011		CELKEM	
		UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč
F1	Osobní Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	1491	1476	2907	2892	2965	2945	2965	2945	2965	2945	0	0	13293	13203
F1.1	Mzdy a platy Mzdy a platy	1066	1066	2100	2100	2150	2150	2150	2150	2150	2150	0	0	9616	9616
F1.2	Dohody Úhrada dohod o pracích konaných mimo pracovní poměr	30	15	30	15	20	0	20	0	20	0	0	0	120	30
F1.3	Povinné zákonné odvody Povinné zákonné odvody včetně připadného přídělu do FKSP	395	395	777	777	795	795	795	795	795	795	0	0	3557	3557
F2	Pořízení majetku Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	1094	640	320	0	0	0	0	0	0	0	0	0	1414	640
F3	Provoz a údržba Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	50	25	100	20	100	0	100	0	100	0	0	0	450	45
F4	Další provozní Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	50	25	100	20	100	0	100	0	80	0	0	0	430	45
F5	Služby Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	10	0	20	0	30	0	30	0	20	0	0	0	110	0
F6	Výsledky Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	10	0	50	0	100	0	100	0	100	0	0	0	360	0
F7	Cestovné Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	100	50	200	65	400	50	400	50	430	50	0	0	1530	265
F8	Doplňkové Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, energii a služby neuvedené výše	150	0	300	0	300	0	300	0	300	0	0	0	1350	0
F9	NÁKLADY CELKEM NÁKLADY CELKEM	2955	2216	3997	2997	3995	2995	3995	2995	3995	2995	0	0	18937	14198
F9A	z toho běžné náklady z toho běžné náklady	1861	1576	3677	2997	3995	2995	3995	2995	3995	2995	0	0	17523	13558

ZDROJE		2006	2007	2008	2009	2010	2011	CELKEM
ZD	Dotace Z dotace na projekt	2216	2997	2995	2995	2995	0	14198
ZO	Ostatní veřejné zdroje Ostatní veřejné zdroje	0	0	0	0	0	0	0
ZN	Neveřejné zdroje Neveřejné zdroje	739	1000	1000	1000	1000	0	4739
ZC	ZDROJE CELKEM ZDROJE CELKEM	2955	3997	3995	3995	3995	0	18937

4.3.2.5. Komentář k uznaným nákladům - Západočeská univerzita v Plzni - Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

Osobní náklady
==========
Mzdové náklady - hlavní položkou mzdových nákladů jsou platy členů řešitelského týmu podle zadaných kritérií. Výše platů je stanovena v souladu se mzdovými předpisy a obvyklými platy u uchazeče. Součástí mzdových nákladů jsou povinné odvody na zdravotní a sociální pojištění v předepsané výši. V 1. roce řešení zahrnují náklady na mzdy a platy řešitelů podle uvedeného rozpisu a náklady na dohody o provedení práce pro anotátory (studenty) a pořizování řečových dat. Dohody o provedení práce budou také pokrývat činnosti, které nebudeme schopni zajistit vlastními silami; půjde především o činnosti spojené s prezentací výsledků centra (výroba rozměrných posterů, audiovizuální zajištění prezentačních akcí, pořádání workshopů aj).
Zvýšení částek na mzdy v následujících letech řešení je způsobeno jednak dvojnásobkem platu každého řešitele oproti 1. roku řešení (v 1. roce řešení se předpokládá pouze 6 měsíců) a jednak zvýšením částek respektujícím nárůst platů v souvislosti s růstem erudice především mladých řešitelů.

Investice (dlouhodobý majetek)
==========
Katedra informatiky a výpočetní techniky plánuje v roce 2006 obnovu serveru Sun Enterprise 250. Server byl zakoupen v roce 2000, v roce 2006 tedy bude vyčerpána doba jeho životnosti. Server v současné době nevyhovuje požadavkům ani z hlediska jeho výkonnosti (dva procesory s taktovací frekvencí 250 MHz) a poskytovaného softwarového komfortu. Předpokládá se proto jeho náhrada serverem o zhruba desetinásobné výkonnosti, umožňujícím instalaci všech současně nabízených softwarových produktů.
Předpokládaná cena zahrnuje DPH; Západočeská univerzita v Plzni je plátcem DPH, ale u projektů výzkumu a vývoje si nenárokuje odpočet.

Provozní náklady
==========
V provozních nákladech je počítáno s pravidelnou údržbou serverového clusteru, zajišťovanou externím dodavatelem, v dalších provozních nákladech je pak počítáno s nákupem běžného spotřebního materiálu, drobného materiálu pro výpočetní techniku, software neinvestičního charakteru, knih apod.

Služby
======
V této kategorii jsou zahrnuty konferenční poplatky a práce prováděné dodavatelsky a hrazené na faktury.

Náklady na zveřejní výsledků
==========
Zahrnují náklady na tisk posterů, vložné na konference, financování výstavních stánků a prezentaci výsledků na speciálních akcích.

Cestovné
========
Cestovné bude použito na cesty na mezinárodní konference (tuzemské i zahraniční), kde budou prezentovány výsledky projektu (průměrně se předpokládá aktivní účast 5 - 7 lidí na konferencích ročně). Z této položky budou též hrazeny cestovní náklady spojené s akcemi přímo vyplývajícími z řešení projektu.

Doplňkové (režijní) náklady
==========
Výše doplňkových (režijních) nákladů, které vzniknou v přímé souvislosti s řešením Centra, je stanovena v souladu s pravidly používanými na ZČU. Prohlašujeme, že pro stanovení výše doplňkových (režijních) nákladů používáme metodu kalkulace dodatečných nákladů a nesplňujeme požadavky na účetní prokazatelnost všech nákladů souvisejících s řešením projektu podle účetní osnovy do maximální výše, stanovené poskytovatelem.

4.3.2. Nákladové tabulky pracovišť
Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

4.3.2.1. Náklady na pořízení majetku - Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

Idč	Název	Dodavatel	Rok pořízení	Pořizovací Cena tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE z SRU tis. Kč	
1	Notebook	Orange&Greeen	2006	75	75	75	
2	Výpočetní server	Orange&Greeen	2007	210	210	210	
	SOUČET ZA ROK		2006		75	75	
	SOUČET ZA ROK		2007		210	210	

4.3.2.2. Náklady na mzdy a platy - Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

	Jméno	Rodné číslo	1. Rok řešení	% pracovního úvazku	měsíční výše platu nebo mzdy včetně nadtarifních složek tis. Kč	počet měsíců	Roční výše odměny tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE z SRU tis. Kč	
	Bártek Luděk Mgr.	720108/3791	2006	50	22	6	5	71	71	
	Horák Aleš Mgr. Ph.D.	740901/4250	2006	20	32	6	10	48,4	48,4	
	Kopeček Ivan doc. RNDr. CSc.	490303/075	2006	50	40	6	10	130	130	
	Pomikálek Jan Mgr.	791009/0419	2006	70	22	6	8	100,4	0	
	Rychlý Pavel Mgr. Ph.D.	730123/5359	2006	50	32	6	10	106	106	
	Sojka Petr RNDr. Ph.D.	630917/1000	2006	50	32	6	10	106	106	
	Pala Karel doc. PhDr. CSc.	390615/416	2006	20	50	6	15	75	75	
	SOUČET		2006					637	537	

4.3.2.3. Náklady na dohody - Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

Idč	Popis činností zajišťovaných v přímé souvislosti s řešením projektu formou dohod o pracích konaných mimo pracovní poměr	1.Rok řešení	Počet hodin	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE z SRU tis. Kč	
1	Tvorba korpusu stromů	2006	200	20	20	
	SOUČET	2006		20	0	

4.3.2.4. Tabulka uznaných nákladů - Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

NÁKLADY		2006		2007		2008		2009		2010		2011		CELKEM	
		UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč	UZNANÉ NÁKLADY tis. Kč	z toho DOTACE ze SRU tis. Kč
F1	Osobní Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	912	755	1746	1472	1822	1521	1850	1542	1877	1562	0	0	8207	6852
F1.1	Mzdy a platy Mzdy a platy	637	537	1274	1074	1330	1110	1350	1125	1370	1140	0	0	5961	4986
F1.2	Dohody Úhrada dohod o pracích konaných mimo pracovní poměr	20	0	0	0	0	0	0	0	0	0	0	0	20	0
F1.3	Povinné zákonné odvody Povinné zákonné odvody včetně případného přídělu do FKSP	255	218	472	398	492	411	500	417	507	422	0	0	2226	1866
F2	Pořízení majetku Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	75	75	210	210	0	0	0	0	0	0	0	0	285	285
F3	Provoz a údržba Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	30	20	60	20	60	40	60	40	60	40	0	0	270	160
F4	Další provozní Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	30	20	60	20	60	40	60	40	60	40	0	0	270	160
F5	Služby Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F6	Výsledky Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F7	Cestovné Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	50	27	120	74	180	121	180	128	180	135	0	0	710	485
F8	Doplňkové Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, energii a služby neuvedené výše	100	0	200	0	200	0	200	0	200	0	0	0	900	0
F9	NÁKLADY CELKEM NÁKLADY CELKEM	1197	897	2396	1796	2322	1722	2350	1750	2377	1777	0	0	10642	7942
F9A	z toho běžné náklady z toho běžné náklady	1122	822	2186	1586	2322	1722	2350	1750	2377	1777	0	0	10357	7657

ZDROJE		2006	2007	2008	2009	2010	2011	CELKEM
ZD	Dotace Z dotace na projekt	897	1796	1722	1750	1777	0	7942
ZO	Ostatní veřejné zdroje Ostatní veřejné zdroje	0	0	0	0	0	0	0
ZN	Neveřejné zdroje Neveřejné zdroje	300	600	600	600	600	0	2700
ZC	ZDROJE CELKEM ZDROJE CELKEM	1197	2396	2322	2350	2377	0	10642

4.3.2.5. Komentář k uznaným nákladům - Masarykova univerzita - Masarykova univerzita Fakulta informatiky Centrum zpracování přirozeného jazyka

Osobní náklady
==========
Mzdové náklady - hlavní položkou mzdových nákladů jsou platy členů řešitelského týmu podle zadaných kritérií. Výše platů je stanovena v souladu se mzdovými předpisy a obvyklými platy u uchazeče. Součástí mzdových nákladů jsou povinné odvody na zdravotní a sociální pojištění v předepsané výši. V 1. roce řešení zahrnují náklady na mzdy a platy řešitelů podle uvedeného rozpisu a náklady na dohody o provedení práce pro lingvisty pracující na přípravě strukturních korpusů. Zvýšení částek na mzdy v následujících letech řešení je způsobeno jednak dvojnásobkem platu každého řešitele oproti 1. roku řešení (v 1. roce řešení se předpokládá pouze 6 měsíců) a jednak zvýšením částek respektujícím nárůst platů v souvislosti s růstem erudice především mladých řešitelů.

Investice (dlouhodobý majetek)
==========
Centrum zpracování přirozeného jazyka plánuje v roce 2006 nákup jednoho notebooku a v roce 2007 nákup výpočetního serveru určeného pro vývoj aplikací a datových korpusů potřebných pro projekt.

Provozní náklady
==========
V provozních nákladech je počítáno s údržbou většinou stávajícího zařízení centra, které bude využito pro práci na projektu. V dalších provozních nákladech je pak počítáno s nákupem běžného spotřebního materiálu, drobného materiálu pro výpočetní techniku, software neinvestičního charakteru, knih apod.

Služby
======
V rámci projektu neplánujeme náklady na služby. Náklady na konferenční poplatky pro řešitele projektu zahrnujeme do části Cestovné.

Cestovné
========
Cestovné bude použito na cesty na mezinárodní konference (tuzemské i zahraniční), kde budou prezentovány výsledky projektu, včetně konferenčních poplatků. Z této položky budou též hrazeny cestovní náklady spojené s akcemi přímo vyplývajícími z řešení projektu.

Doplňkové (režijní) náklady
==========
Výše doplňkových (režijních) nákladů, které vzniknou v přímé souvislosti s řešením projektu, je stanovena v souladu s pravidly používanými na Masarykově univerzitě.

5.1. Údaje pro IS VaV

Název projektu anglicky
Complex knowledge base tools for natural language communication with the semantic web

Cíl předmětu řešení anglicky
The aim is to design a complex system of formal and implementation tools for building user-friendly interfaces to semantic web. These tools are based on principles of intelligent agents and enable to communicate with user in natural language. Then the proceesed data have also a form of sentences (utterances) of natural language. An additional aim includes verification of functional attributes of proposed tools on a specific application.

Klíčová slova česky
Interakce člověk - počítač, inteligentní komunikační rozhraní, komunikace v přirozeném jayzce, znalostní systémy a znalostní inženýrství, sémantický web

Klíčová slova anglicky
Human - computer interaction, intelligent communication interface, natural language communication, knowledge systems and knowledge engineering, semantic Web

Klasifikace hlavního oboru řešení
JC

Klasifikace vedlejšího oboru řešení
JD

Klasifikace dalšího vedlejšího oboru řešení
AI

Stupeň důvěrnosti údajů
S

Kategorie výzkumu a vývoje
Aplikovaný výzkum s výjimkou průmyslového výzkumu (tzv. "neprůmyslový výzkum")

Vyjádření ke stanovení podílu účelové podpory
Pracoviště budou hradit 25% nákladů z vlastních neveřejných zdrojů, výzkumné činnosti při řešení projektu nebudou obsahovat průmyslový výzkum a vývoj.

5.2. Osoby, které by se mohly vyjádřit k návrhu

Krokavec Dušan prof. Ing. CSc. Slovenská republika
Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
++421 55 6253564 ++421 55 6253574 Dusan.Krokavec@tuke.sk
Pokorný Jaroslav prof. RNDr. CSc. 480423/016 Česká republika
0021620 Univerzita Karlova v Praze
11320 Matematicko-fyzikální fakulta
221 914 265 221 914 323 pokorny@ksi.ms.mff.cuni.cz
Novák Mirko prof. Ing. DrSc. Česká republika
68407700 České vysoké učení technické v Praze
21260 Fakulta dopravní
224 359 548 224 359 545 novak@fd.cvut.cz
Slavík Pavel prof. Ing. CSc. 460415/073 Česká republika
68407700 České vysoké učení technické v Praze
21230 Fakulta elektrotechnická
224 357 617 224 923 325 slavik@cslab.felk.cvut.cz

5.3. Osoby, kterým by se neměl předkládat návrh projektu k vyjádření

Nouza Jan prof. Ing. CSc. Česká republika
46747885 Technická univerzita v Liberci
24220 Fakulta mechatroniky a mezioborových inženýrských studií
485 353 208 485 353 112 jan.nouza@tul.cz